Universidade de Aveiro
2008

Departamento de Electrónica,
Telecomunicações e Informática

João Tiago da Rocha Araújo

Impacto do scaling da tecnologia CMOS no desenho de circuitos digitais

Universidade de Aveiro
2008

Departamento de Electrónica,
Telecomunicações e Informática

João Tiago da Rocha Araújo

Impacto do scaling da tecnologia CMOS no desenho de circuitos digitais

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Electrónica e Telecomunicações, realizada sob a orientação científica do Doutor Ernesto Fernando Ventura Martins, Professor Auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro, e do Mestre Luís Filipe Mesquita Nero Moreira Alves, Professor Assistente do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro.

Aos meus pais e ao meu irmão.
À Vânia.
A todos os meus familiares e amigos.

o júri

presidente

Professor Doutor Dinis Gomes de Magalhães dos Santos
Professor Catedrático do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

Professor Doutor João Carlos da Palma Goes
Professor Auxiliar do Departamento de Engenharia Electrotécnica da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

Professor Doutor Ernesto Fernando Ventura Martins
Professor Auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

agradecimentos

As minhas primeiras palavras são para os meus pais, Alberto José e Maria Rosa, aos quais agradeço por tudo o que sou hoje, pelas condições que me proporcionaram e pelo apoio incondicional que sempre me deram, não tanto durante este trabalho, mas principalmente ao longo de toda a minha vida e de todo o meu percurso escolar e académico. Ao meu irmão, Fernando Jorge, agradeço também pelos momentos de lazer, pela companhia, pelas discussões, por tudo. Essencialmente por ser também um bom irmão.

Agradeço à Vânia, a minha namorada. A minha melhor amiga e o meu porto de abrigo de sempre. Confidente, companheira nos bons e nos maus momentos, a ela agradeço toda a compreensão, todo o cuidado e todo o carinho demonstrados, mesmo quando por este ou por aquele motivo sacrifiquei algumas horas só nossas. Mostro-me grato também à minha restante família, por todos os motivos e mais alguns, que agora não me vale a pena especificar.

Já não era sem tempo: aos meus orientadores, o Professor Doutor Ernesto Ventura Martins e o Professor Mestre Luís Nero Alves. Pela confiança no meu trabalho demonstrada desde a primeira hora, pelos conhecimentos transmitidos, pelo rigor, pela atitude crítica das situações menos consensuais, pela organização e método de trabalho e, fundamentalmente, pelo acompanhamento ímpar que fizeram do estudo que fui realizando.

Grato às instituições onde dispus das melhores condições para cumprir o Ensino Superior e para este trabalho de dissertação: a Universidade de Aveiro, o Departamento de Electrónica, Telecomunicações e Informática e o Instituto de Telecomunicações – Pólo de Aveiro.

Por último, como não poderia deixar de ser, a todos os colegas que me acompanharam durante o meu percurso académico. Cada um no seu momento, todos me marcaram à sua maneira pelo convívio, amizade e companheirismo. Muitos ficarão como amigos. De outros espero que não se perca o contacto. No entanto, Aveiro será sempre nossa.

A todos (e aqui incluem-se também os que me esqueci), o meu sincero obrigado.

João Tiago da Rocha Araújo

palavras-chave

tecnologia CMOS, scaling, técnicas de desenho.

resumo

Este trabalho de dissertação insere-se na área da electrónica digital e visa avaliar as técnicas tradicionais de desenho de circuitos CMOS. O rápido desenvolvimento das tecnologias CMOS, sustentado pelas teorias de *scaling*, tem vindo a suscitar o interesse na criação de novos modelos analíticos e a proporcionar vários desafios ao nível do projecto de circuitos digitais.

A principal motivação deste trabalho prende-se, por isso mesmo, com o estudo do impacto do *scaling* no desenho e na optimização de circuitos das tecnologias actuais. As técnicas convencionais de desenho foram formuladas há algumas décadas atrás, pelo que a constante redução das dimensões dos dispositivos tem revelado a ineficácia destas mesmas técnicas aplicadas ao projecto de portas lógicas das tecnologias correntes.

Deste modo, este trabalho foca-se nalguns desses desafios inerentes ao desenho optimizado de circuitos que utilizem transístores de canal curto. Pretende-se um estudo relativamente amplo, pelo que se propõe a caracterização de diversas portas lógicas CMOS estáticas, utilizando no decorrer do plano de trabalhos cinco tecnologias diferentes. O desenho das portas lógicas é feito no ambiente integrado do *Cadence*, enquanto o trabalho de caracterização utiliza o simulador *Spectre*.

keywords CMOS technology, scaling, design techniques.

abstract

This thesis presents aspects that are related with the digital electronic design area, and aims to evaluate the traditional design techniques of CMOS circuits. The sudden development of CMOS technology, supported by scaling theories, has already led to the interest in creating new analytical models, and simultaneously has posed various challenges in the design of digital circuits.

The main contribution of this thesis is the study of the impact of scaling in the design and optimization of digital circuits in current CMOS technologies. The conventional design techniques were advanced a few decades ago, hence the constant reduction shows that these techniques are no longer appropriate for the project of logic gates optimized, for the current technologies.

Therefore, all the work related with this thesis could not avoid some challenges associated with the design of optimized circuits with short-channel devices. It is a relatively wide study, so the characterization of static CMOS logic gates is done recurring to five different technologies along the planning of work. The design of the logic gates is made with the *Cadence* tools, while the work of characterization of these gates uses the simulator *Spectre*.

“it's easy to look at a neuron and a transistor and say that one is slow and one is fast, but the mind is harder to understand”

The Singularity Institute for Artificial Intelligence

Índice

1 Introdução 1

1.1 Impacto do *scaling* nas tecnologias actuais 3

1.2 Objectivos e motivação 4

1.3 Metodologia 4

1.4 Estrutura da dissertação 5

2 Fundamentos sobre tecnologias CMOS 7

2.1 O transístor MOSFET 7

2.1.1 Física do dispositivo 8

2.1.2 Modelo de funcionamento do MOSFET 9

2.1.3 Efeitos de segunda ordem 11

2.1.3.1 Corrente de sublimiar 11

2.1.3.2 Efeito de corpo 12

2.1.3.3 Modulação do comprimento de canal 13

2.1.3.4 Variações da tensão de limiar V_{TH} 14

2.1.3.5 Saturação de velocidade 14

2.1.3.6 Degradação da mobilidade 16

2.2 Capacidades do MOSFET 17

2.2.1 Capacidades da porta para o canal 18

2.2.2 Capacidades de *overlap* 18

2.2.3 Capacidades das junções PN 19

2.2.4 Modelo de capacidades do MOSFET 20

2.2.5 Comportamento das ligações 21

2.3 Teoria de *Scaling* 23

2.3.1 *Scaling* de campo eléctrico constante 23

2.3.2 *Scaling* de tensões fixas 24

2.3.3 *Scaling* geral 25

2.3.4 Impacto do *scaling* na física dos dispositivos MOS 26

2.4 Sumário 29

3 Desenho de Circuitos CMOS Estáticos 31

3.1 O inversor CMOS 31

3.2 Técnicas de desenho tradicionais 34

3.2.1 Influência das diferenças entre tensões de limiar 35

3.2.2 Influência da capacidade de carga e da transição de entrada 36

3.2.3 Influência da capacidade de Miller 37

3.3 Circuitos Combinatórios 38

3.3.1 Construção do modelo do inversor equivalente 40

3.4 Sumário 43

4 Caracterização das portas lógicas 45

4.1 Tecnologias CMOS utilizadas 45

4.2 Método de desenvolvimento 46

4.3 Estratégias utilizadas no desenho das portas lógicas 48

4.4 Resultados 50

4.4.1 Caracterização do inversor estático 50

4.4.1.1 Estudo isolado de uma tecnologia 50

4.4.1.2 Comparação entre as cinco tecnologias utilizadas 56

4.4.2 Caracterização das portas NAND e NOR 59

4.5 Sumário 61

5 Conclusões 63

5.1 Linhas de investigação futuras 64

Referências 67

Lista de Figuras

2.1 Símbolos do MOSFET: (a) do tipo N; (b) do tipo P. 8

2.2 Vista em forte do MOSFET do tipo N. 8

2.3 Família de curvas do MOSFET: característica I-V tendo V_{GS} como parâmetro. 10

2.4 Vista em corte do MOSFET do tipo N na região de saturação. 11

2.5 Modulação do comprimento de canal: característica I-V do MOSFET. 13

2.6 Saturação da velocidade de deriva. 15

2.7 Capacidades do MOSFET. 17

2.8 Capacidades da porta para o canal do MOSFET. 18

2.9 Capacidades de sobreposição da porta com as regiões da fonte e do dreno. 18

2.10 Vista em detalhe da junção PN da fonte do MOSFET. 19

2.11 Modelo de capacidades do MOSFET. 20

2.12 Estrutura tridimensional de uma ligação. 21

2.13 Vista em detalhe das camadas de ligações entre dispositivos. 21

2.14 Princípios do *scaling*: dispositivo original (a) e reduzido (b). 23

2.15 Transístores de canais longo e curto: comparação entre características I-V. 26

2.16 Estruturas SOI: (a) GAA MOSFET e (b) FinFET. 28

3.1 O Inversor CMOS. 32

3.2 Inversor durante a carga. 32

3.3 Inversor durante a descarga. 33

3.4 Modelo simplificado do inversor para entrada alta (b) e baixa (c). 34

3.5 Inversor CMOS com capacidade Miller entre a entrada e a saída. 37

3.6 Constituição de um circuito combinatório. 39

3.7 Porta NOR duas entradas. 40

3.8 Porta NAND duas entradas. 42

4.1 Utilização do Cadence como ferramenta desenvolvimento. 47

4.2 Vista de desenho *layout*. 48

4.3 Esquemático do ambiente de teste de um inversor estático. 49

4.4 UMC130: desequilíbrio δ em função do tamanho do inversor para a técnica da razão das mobilidades. 51

4.5 UMC130: influência da razão de desenho. 53

4.6 UMC130: desequilíbrio δ em função do *fan-out*. 54
4.7 UMC130: influência do *fan-out* e do *trin* no desenho do INV8X. 54
4.8 AMS800: influência do *fan-out* e do *trin* no desenho do INV8X. 55
4.9 Caracterização das diferentes tecnologias em termos de desequilíbrio δ. 56
4.10 Caracterização das diferentes tecnologias em termos de desequilíbrio δ para o ajuste de desenho empírico. 58

Lista de Tabelas

1.1	Capacidades do MOSFET dependentes do canal de inversão.	19
1.2	Influência do *scaling* nos diferentes parâmetros do MOSFET.	26
4.1	Análise dos piores casos das portas da tecnologia UMC130.	59
4.2	Análise dos piores casos das portas da tecnologia AMS800.	61

Lista de Símbolos

β	razão de desenho
C_{DB}	capacidade dreno-substrato
C_{fringe}	capacidade lateral
C_G	capacidade intrínseca da porta
C_{GC}	capacidade da porta para o canal do MOSFET
C_{GCB}	capacidade porta-substrato
C_{GCD}	capacidade porta-dreno
C_{GCS}	capacidade porta-fonte
C_{GDO}	capacidade de sobreposição porta-dreno
C_{GSO}	capacidade de sobreposição porta-fonte
C_{inter}	capacidade entre fios
C_j	capacidade da junção da base
C_{jsw}	capacidade da junção lateral
C_L	capacidade de carga
C_M	capacidade de Miller
C_o	capacidade por unidade de largura
C_{ox}	capacidade porta-substrato por unidade de área
C_{pp}	capacidade do condensador de placas paralelas
C_{SB}	capacidade fonte-substrato
C_{wire}	capacidade das ligações
δ	desequilíbrio entre os tempos de propagação
ξ_c	campo eléctrico crítico (valor limiar do campo eléctrico)
ε_{di}	permitividade do dieléctrico
ξ_x	campo eléctrico
H	espessura das ligações de um circuito
$i_{D,n}$	corrente de dreno do NMOS
$i_{D,p}$	corrente de dreno do PMOS
I_{DS}	corrente de dreno
Φ_F	potencial de Fermi
k'	parâmetro de transcondutância

k	transcondutância do MOSFET
k_B	constante de Boltzmann
λ	factor de modulação do comprimento de canal
L	comprimento do canal de um transístor
L_{eff}	comprimento efectivo do canal de um transístor
q	carga do electrão
R_{eqN}	resistência equivalente do transístor de tipo N
R_{eqP}	resistência equivalente do transístor de tipo P
T	temperatura
t_{di}	espessura do dieléctrico
t_{ox}	espessura do óxido da porta (óxido fino)
t_p	tempo (atraso) de propagação
t_{pHL}	tempo de propagação *high-to-low*
t_{pLH}	tempo de propagação *low-to-high*
μ	mobilidade dos portadores
v_d	velocidade de deriva
V_{DD}	tensão de alimentação positiva
V_{DS}	tensão dreno-fonte
V_{GD}	tensão porta-dreno
V_{GS}	tensão porta-fonte
V_{in}	tensão de entrada
V_M	tensão de comutação do inversor
V_{out}	tensão de saída
V_{SB}	tensão fonte-substrato
V_{TH}	tensão de limiar
V_{TH0}	tensão de limiar para $V_{SB} = 0$
W	largura do canal de um transístor
x_d	difusão lateral
X_j	profundidade da junção
γ	factor de efeito de corpo

Lista de Acrónimos

CMOS	*Complementary Metal-Oxide Semiconductor*
ADE	*Analog Design Environment*
AMS	*Austria Microsystems*
BOX	*Buried Oxide*
BSIM	*Berkeley Short-channel IGFET Model*
CAD	*Computer-Aided Design*
DF	*Design Framework*
DG	*Double-Gate*
DGSOI	*Double-Gate Silicon-On-Insulator*
DIBL	*Drain-Induced Barrier Lowering*
DRC	*Design Rule Check*
DTMOS	*Dynamic Threshold MOS*
FDK	*Foundry Design Kit*
FET	*Field-Effect Transistor*
GAA	*Gate-All-Around*
ITRS	*International Technology Roadmap for Semiconductors*
LVS	*Layout Versus Schematic*
MOS	*Metal-Oxide Semiconductor*
MOSFET	*Metal-Oxide Semiconductor Field-Effect Transistor*
PDN	*Pull-Down Network*
PUN	*Pull-Up Network*
RCX	*Resistance/Capacitance and Inductance Extraction*
SOI	*Silicon-On-Insulator*
UMC	*United Microelectronics Corporation*
VLSI	*Very Large Scale Integration*
VTC	*Voltage Transfer Characteristic*

Capítulo 1

Introdução

O sucesso da indústria dos semicondutores na última metade do século XX teve, embora sem o devido reconhecimento por parte da sociedade, grande relevância nas nossas vidas. Dos computadores às mais diversas aplicações domésticas, passando pela indústria automóvel, pela medicina ou pelo mundo das telecomunicações móveis, tudo o que de alta tecnologia nos rodeia hoje em dia provém do grande crescimento que o sector dos circuitos integrados viveu nas últimas décadas. Os avanços notáveis ao nível da capacidade de integração de circuitos quebraram todas as barreiras do que era expectável e tiveram um papel muito importante na própria sociedade.

O primeiro circuito integrado foi desenvolvido no ano de 1958, em separado por Jack Kilby, da *Texas Instruments*, e Robert Noyce, da *Fairchild Semiconductor* [1, 2, 25]. Kilby fê-lo numa placa de germânio e Noyce, poucos meses depois, deixou o germânio de lado e utilizou silício como material semicondutor base. A invenção, da qual os dois são tidos como co-inventores, revolucionou por completo a indústria electrónica e, poucos anos mais tarde [3], proporcionou o desenvolvimento da tecnologia CMOS (*Complementary Metal-Oxide Semiconductor*), que viria a revelar-se fundamental na evolução das aplicações VLSI (*Very Large Scale Integration*). Contudo, foi a previsão de Gordon Moore, feita em 1965, que sustentou o grande desenvolvimento das capacidades do circuito integrado. A observação empírica do co-fundador da *Intel Corporation*, que dizia que a capacidade de integração de transístores num só *chip* iria duplicar a cada dois

anos, ficou conhecida por Lei de Moore [4] e, espantosamente, mantém-se válida na actualidade (apesar das pequenas correcções a que foi sujeita). Moore previa que o circuito integrado atingisse patamares inimagináveis na altura, vaticinando a sua produção em massa. Não se enganou, uma vez que ainda nos dias de hoje o seu vaticínio é tido em consideração por toda a indústria dos circuitos integrados [28, 37].

No entanto, para que a Lei de Moore tivesse suporte, era necessário reduzir as dimensões físicas dos transístores de forma a garantir uma maior capacidade de integração dos mesmos no mesmo circuito. Para o grande desenvolvimento dos circuitos, em muito contribuiu a evolução dos processos litográficos (que estão na génese do circuito integrado). E foi no sentido de aumentar a escala de integração que, no início da década de 70, se formularam as primeiras teorias de *scaling*. Robert Dennard foi quem introduziu o tema pela primeira vez [7], demonstrando que se todas as dimensões de um transístor MOS forem reduzidas simultaneamente, juntamente com alterações proporcionais nas tensões de funcionamento [8], podiam-se continuar a fabricar dispositivos cada vez mais pequenos, preservando as suas características eléctricas e operacionais. Ao formular este modelo que possibilitava o fabrico de transístores mais pequenos e com melhores desempenhos (mais rápidos e densos, com menor consumo de potência, tudo com uma diminuição nos custos de produção por *chip*), Dennard providenciou um meio para sustentar a Lei de Moore.

O progressivo *scaling* da tecnologia levou a que, no espaço de quatro décadas, já se estivessem a desenvolver transístores abaixo da ordem dos micrometros. A tecnologia, entretanto, para continuar a seguir a Lei de Moore foi sendo forçada a adoptar novas estratégias. Por isso mesmo, a *Intel* desenvolveu nos últimos anos uma inovadora tecnologia de fabricação juntamente com os seus minúsculos transístores de 45 nm: a denominada "High-K e Metal Gate" [39]. Trata-se de uma combinação de novos materiais com uma propriedade designada por "high-k" – material isolante com características dieléctricas propícias ao fabrico da porta do transístor – que utiliza ainda uma outra mescla de materiais metálicos na construção da porta propriamente dita – usa o metal de transição háfnio em vez de recorrer ao polisilício [39].

A família de processadores "Penryn", lançada pela *Intel* em 2007, já adoptou esse processo litográfico de 45 nm. O ano de 2007 marcou também a demonstração, em funcionamento, do primeiro *chip* da indústria fabricado com a tecnologia de 32 nm e a litografia de 32 nm está prevista para começar a ser utilizada dentro de um ano, em 2009.

Nos próximos dez anos, de acordo com o ITRS (*International Technology Roadmap for Semiconductors*), o processo tecnológico de 32 nm terá como sucessores o de 22 nm (previsto para 2012) e o de 16 nm (previsto para 2018). O que reflecte uma tendência que se vem estabelecendo ao longo do tempo: o *scaling* da tecnologia leva a que, de geração para geração, as dimensões dos transístores sejam reduzidas em cerca de 70%. A base deste crescimento mais sustentado deve-se ao facto das últimas gerações tecnológicas terem sido planeadas convenientemente. Conseguiu-se um aumento de desempenho ao privilegiar uma política diferente: os fabricantes passaram a investigar e a dedicarem-se a mais assuntos ao mesmo tempo, ao invés de optarem por uma busca desenfreada de novas soluções e novas gerações de transístores, como vaticinava originalmente a Lei de Moore.

1.1 – Impacto do *scaling* nas tecnologias actuais

À medida que a tecnologia avança para dispositivos de canal cada vez mais curto (inferior a 100 nm), escusado será dizer que os desenhadores de circuitos enfrentam cada vez mais desafios. A necessidade de integrar cada vez mais transístores numa área reduzida trouxe grandes vantagens, como a maior densidade dos circuitos ou maior rapidez dos dispositivos. Todavia, a redução das dimensões físicas dos dispositivos levou a que efeitos secundários se tornassem mais intensos e passassem a influenciar o comportamento dos transístores. Ganharam, portanto, extrema relevância os efeitos de canal curto.

Os efeitos de canal curto assumiram então grande importância nos transístores actuais. A saturação de velocidade de deriva dos portadores, assim como a degradação da mobilidade dos mesmos, motivados pelos intensos campos eléctricos a que estão sujeitos estes dispositivos, foram dois dos fenómenos mais importantes a este nível. Mas outros surgiram como consequência do *scaling*. A condução na região sublimiar é outro dos efeitos de canal curto e trouxe grandes implicações ao nível do consumo estático, principalmente devido ao acentuado aumento das correntes de fuga, que são seguramente o maior desafio que a tecnologia enfrenta nos transístores com canais curtos, para dimensões abaixo dos 100nm [24, 28].

À rápida cadência do *scaling* e ao agravamento dos efeitos de canal curto atribui-se mesmo a supressão de algumas considerações tidas para o desenho de circuitos digitais. A evolução da tecnologia levou a que as técnicas tradicionais de desenho deixassem de

possibilitar o projecto de circuitos digitais optimizados como outrora, à medida que caminha para dispositivos de dimensões cada vez mais reduzidas. É neste ponto, no estudo do impacto do *scaling* nessas técnicas tradicionais, que se foca este trabalho de dissertação.

1.2 – Objectivos e motivação

De acordo com os argumentos apresentados anteriormente, a evolução tecnológica, aliada ao desenvolvimento das teorias de *scaling*, levou a que na actualidade só se fale em transístores de canal curto. Contudo, a constante redução das dimensões dos dispositivos originou uma série de dificuldades incontornáveis no cumprimento dos requisitos de operação dos transístores, dificuldades essas que implicam novos desafios ao nível do desenho de circuitos.

O estudo dessas barreiras impostas pelo *scaling* é por isso o objectivo principal deste trabalho, uma vez que as técnicas tradicionais de desenho deixaram de ser apropriadas para projectar circuitos optimizados como outrora. Neste trabalho, pretende-se oferecer uma perspectiva diferente do que é trabalhar com transístores de dimensões reduzidas e inferir linhas de orientação que possibilitem um projecto optimizado e uma maior eficiência destes circuitos, dando uma contribuição própria para o tema.

1.3 – Metodologia

O objectivo fundamental desta dissertação prende-se com a avaliação do impacto do *scaling* no desenho de circuitos nas tecnologias actuais, de canal curto. Assim, será dada maior importância ao estudo das tecnologias de 180 nm, 130 nm e 90 nm. Contudo, como se pretende uma análise mais abrangente, as tecnologias de 800 nm e 350 nm são também alvo de estudo. Deste modo, do plano de trabalhos faz parte o desenho e caracterização de uma série de portas lógicas, utilizando para o efeito os *design-kits* providenciados pela UMC (*United Microelectronics Corporation*) e pela AMS (*Austria Microsystems*), ao abrigo do protocolo firmado entre a Universidade de Aveiro e o *Europractice*[1].

[1] **Europractice** – o *Europractice* oferece condições especiais às universidades e aos institutos de investigação europeus, disponibilizando, através de um contrato estabelecido com a Comissão Europeia, *design-kits* para uso académico, o acesso a ferramentas de CAD e descontos especiais no fabrico de protótipos.

O desenho das portas lógicas, quer de esquemático, quer de *layout*, foi desenvolvido dentro do ambiente integrado do *Cadence DFII* (*Design Framework II*), ambiente esse configurado, para cada tecnologia, com o respectivo *design-kit* fornecido pelo *Europractice* para fins académicos. A caracterização das portas lógicas desenhadas foi feita através de diversos testes e simulações realizados também em ambiente do *Cadence*, mas utilizando o simulador *Spectre* e a sua interface com o utilizador, o *Analog Design Environment* (ADE).

1.4 – Estrutura da dissertação

Esta dissertação encontra-se organizada em cinco capítulos. No Capítulo 1 é introduzido o tema do trabalho, com uma abordagem histórica que vai desde os primórdios do circuito integrado até à introdução do processo litográfico mais recente, ligando as duas épocas e ressalvando a importância da Lei de Moore nas várias gerações tecnológicas e no surgimento das primeiras teorias de *scaling*. É dado também ênfase às perspectivas futuras da tecnologia e referem-se as motivações e objectivos do trabalho.

O Capítulo 2 centra-se na apresentação dos fundamentos relativos às tecnologias CMOS. Para a elaboração de qualquer estudo sobre o comportamento de um dado circuito, é importante dominar o comportamento dos dispositivos que o compõem, pelo que o capítulo começa por concentrar atenções no modelo de funcionamento do transístor MOSFET e das suas características físicas. No mesmo âmbito, é introduzida a teoria de *scaling* e as suas diferentes concepções, referindo-se o impacto que a progressiva redução das dimensões dos dispositivos tem na física e na teoria convencional do MOSFET.

O desenho de circuitos CMOS estáticos é o assunto essencial do Capítulo 3. Começa-se por apresentar o inversor estático (que será sempre o ponto de partida e de contacto para todos os estudos realizados ao longo deste trabalho) e abordar as suas características mais importantes para o desenho de circuitos. Depois, são discutidas as técnicas tradicionais utilizadas no desenho de portas lógicas estáticas, bem como as derivações a esse modelo convencional propostas por alguns autores nos anos mais recentes. A teoria de desenho de circuitos combinatórios, que possibilita a extensão da análise feita para o inversor para outras portas lógicas com diversas entradas, é também apresentada nas derradeiras secções deste capítulo.

No Capítulo 4, expõem-se os resultados do trabalho de caracterização das portas lógicas desenhadas. São descritas as tecnologias CMOS utilizadas neste estudo, bem como a metodologia de desenvolvimento do trabalho e as estratégias escolhidas ao longo do mesmo. Enunciam-se as simulações realizadas e interpretam-se os resultados obtidos para cada tecnologia alvo de estudo.

Por último, o Capítulo 5 resume toda a dissertação, apresentando as conclusões e discutindo os resultados mais importantes que se retiram de todo o trabalho desenvolvido. É igualmente referida a contribuição do estudo para o tema e são detalhadas algumas linhas de investigação futuras, de modo a dar seguimento ao estudo efectuado.

Capítulo 2

Fundamentos sobre tecnologias CMOS

Neste capítulo pretende-se explorar os conceitos de base das tecnologias CMOS. A tecnologia complementar CMOS utiliza quer transístores do tipo P quer transístores do tipo N e foi apresentada por Frank Wanlass e Chih-Tang Sah [3], em 1963. Na altura surgiu como alternativa aos processos que apenas utilizavam transístores NMOS e rapidamente foi ganhando-lhes terreno, devido à inerente redução do consumo de potência dos circuitos assim construídos. Actualmente, é a tecnologia dominante ao nível do desenho de circuitos. Nas secções que se seguem, começa-se por abordar a teoria convencional do transístor, o seu modelo de operação e as suas principais características. É dado relevo aos principais efeitos secundários que afectam o funcionamento do transístor, especialmente para dispositivos de canal curto. O capítulo fecha-se com a introdução do conceito de *scaling* e os seus diversos modelos, realçando-se o impacto que a constante redução das dimensões tem ao nível da física dos dispositivos.

2.1 – O transístor MOSFET

O transístor MOSFET (*Metal-Oxide Semiconductor Field-Effect Transistor*) é o dispositivo base de mais de 95% dos circuitos integrados digitais da actualidade. Transístor de fácil fabrico – o que significa desde logo baixos custos – o MOSFET caracteriza-se pelo baixo consumo estático e pela reduzida área de silício ocupada por

cada transístor [10]. Os seus símbolos são apresentados na Figura 2.1, com os seus quatro terminais *G*, *S*, *D* e *B* a designarem-se, respectivamente, por porta (*gate*), fonte (*source*), dreno (*drain*) e substrato (*body*).

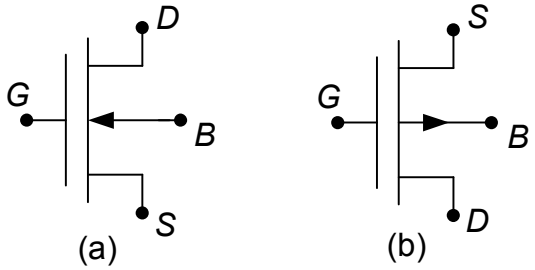

Figura 2.1 – Símbolos do MOSFET: (a) do tipo N; (b) do tipo P.

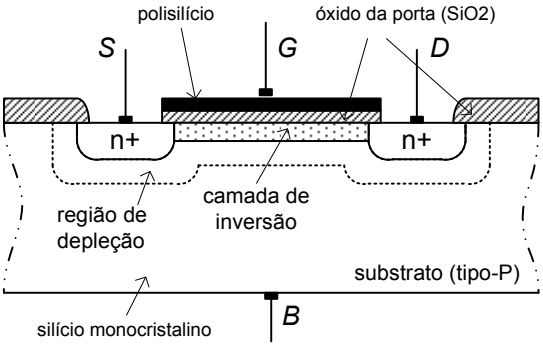

Figura 2.2 – Vista em corte do MOSFET do tipo N.

Existem dois tipos de MOSFET, complementares entre si: os NMOS, cujo substrato é do tipo P e cuja condução é assegurada pelos electrões; e os PMOS, com substrato do tipo N e condução assegurada por lacunas [10]. A Figura 2.2 mostra a vista em corte de um MOSFET do tipo N, sendo possível identificar a estrutura do canal, composta pela porta (polisilício), por um material isolante (dióxido de silício, SiO_2, o chamado óxido da porta) e pelo substrato (silício monocristalino).

2.1.1 – Física do dispositivo

No MOSFET do tipo N, as regiões da fonte e do dreno são similares, como se pode ver na Figura 2.2. São as tensões aplicadas aos terminais do dispositivo que determinam qual a região n^+ que fornece electrões (identificada com o terminal de fonte) e qual a região que recebe esses electrões (terminal de dreno). As tensões aplicadas à porta e ao dreno, assim como a aplicada ao substrato, têm sempre como referência o potencial da fonte [10]. Daí que venham expressas como V_{GS}, V_{DS} e V_{BS}.

Resumidamente, como o próprio acrónimo indica, o MOSFET trabalha sob o princípio de modificar o campo eléctrico do substrato situado por baixo da porta [7]. O campo eléctrico num transístor é provocado por uma diferença de potencial e tem duas componentes: uma componente longitudinal, na direcção do comprimento do canal; e uma componente transversal, perpendicular à interface óxido-semicondutor [15]. No MOSFET,

a condução de corrente é sempre assegurada por um tipo de portadores de carga, existindo dois mecanismos responsáveis pelo movimento de portadores: a deriva e a difusão. A acção de um campo eléctrico externo acelera os portadores, sendo este mecanismo de condução denominado de corrente de deriva. O processo de deriva é responsável pela inversão forte do canal. Um outro mecanismo diferente designa-se por difusão e é causado pelo gradiente não nulo de concentração de portadores de carga. Quando uma região de tipo P é colocada em contacto com uma região de tipo N para formar uma junção, como a concentração de electrões numa região é mais elevada do que na outra região, ocorre a difusão de electrões para dentro da região de menor concentração [10]. Este processo origina uma corrente de difusão.

Na Figura 2.2 podem ver-se ainda duas regiões muito importantes num MOSFET, a região de depleção e a camada de inversão. O modo de operação do transístor está relacionado com o grau de inversão do canal [10]. Aplicando uma tensão positiva na porta (com a fonte e o dreno ligados à massa), leva-se a que as cargas positivas situadas debaixo da porta sejam repelidas, formando uma região esvaziada de portadores (dita de depleção) e dando origem a uma acumulação de electrões minoritários na mesma zona, o que está na base da inversão do canal. A tensão positiva da porta faz com que os electrões das regiões n^+ da fonte e do dreno sejam atraídos para o canal por baixo da porta, induzindo-se um canal *n* que liga as regiões da fonte e do dreno. Aplicando agora uma tensão positiva entre o dreno e a fonte, passa a fluir uma corrente nesse canal induzido. Como esse canal é induzido invertendo a superfície do substrato do tipo *p* para o tipo *n*, designa-se o canal induzido por camada de inversão [10].

Quanto ao grau de inversão do canal, pode-se estar perante inversão fraca (quando ainda não existe canal e a região do substrato por baixo da porta está fracamente invertida), forte (quando existe canal induzido) e moderada (transição entre a inversão fraca e forte). É esse grau de inversão que distingue os modos de funcionamento do transístor.

2.1.2 – Modelo de funcionamento do MOSFET

Para derivar o modelo de funcionamento do MOSFET, fundamenta-se que o canal de inversão de cargas depende da tensão aplicada na porta a partir de um determinado limiar, designado por V_{TH}, a tensão de limiar do transístor. Enquanto que abaixo desse limiar não existe camada de inversão (apesar de fluir uma corrente de sublimiar), diz-se que a partir

desse limiar V_{TH} ocorre a inversão forte do canal. Assim, por outras palavras, é a tensão aplicada à porta que controla o fluxo de electrões da fonte para o dreno. Num transístor, quanto ao seu modelo de funcionamento (Figura 2.3), distinguem-se três regimes de funcionamento: corte, linear e saturação [10, 15].

Na região de corte, isto é, para $V_{GS} < V_{TH}$, a corrente entre o dreno e a fonte, I_{DS}, é essencialmente nula, uma vez que o canal induzido ainda não atingiu a inversão desejada. No entanto, logo que a tensão V_{GS} atinge a tensão de limiar V_{TH}, ocorre a inversão forte, formando-se um canal de condução de cargas que liga as regiões de dreno e da fonte. Ou seja, para $V_{GS} > V_{TH}$, desde que $V_{DS} \neq 0$, há condução e o MOSFET passa a operar na zona linear, primeiro, e depois na região de saturação. Para $V_{DS} < V_{GS} - V_{TH}$, o transístor opera no regime linear e a sua característica tensão-corrente rege-se pela seguinte equação, do conhecido modelo quadrático de Harold Shichman e David Hodges (modelo Shichman-Hodges) [5]

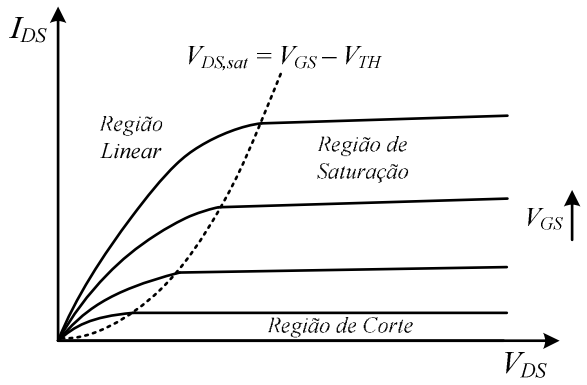

Figura 2.3 – Família de curvas do MOSFET: característica I-V tendo V_{GS} como parâmetro.

$$I_{DS} = k_n' \frac{W}{L}\left[(V_{GS} - V_{TH})\, V_{DS} - \frac{V_{DS}^{\,2}}{2}\right] = k_n \left[(V_{GS} - V_{TH})\, V_{DS} - \frac{V_{DS}^{\,2}}{2}\right] \qquad (2.1)$$

com W (largura) e L (comprimento) como dimensões do canal do transístor e k_n', conhecido por parâmetro de transcondutância do processo, uma constante expressa por

$$k_n' = \mu_n\, C_{ox} = \mu_n \frac{\varepsilon_{ox}}{t_{ox}} \qquad (2.2)$$

onde μ_n é a mobilidade superficial dos electrões, C_{ox} a capacidade por unidade de área do óxido da porta, também conhecida por capacidade porta-substrato, ε_{ox} a permitividade e t_{ox} a espessura do óxido da porta. Para pequenos valores de V_{DS}, o factor quadrático da expressão (2.1) pode ser negligenciado e o comportamento do transístor assemelha-se ao de uma resistência, pois observa-se uma dependência aproximadamente linear entre V_{DS} e I_{DS}. Daí que por vezes também se designe esta região por região óhmica, na qual o transístor, num modelo bastante simplificado, actua como uma resistência controlada pela diferença entre V_{GS} e V_{TH}.

Noutras circunstâncias, para valores superiores da tensão V_{DS}, a corrente I_{DS} deixa de aumentar, estabiliza e aproxima-se de um valor constante. Nesse caso, diz-se que a corrente satura, porque o canal de inversão vai diminuído de espessura e acaba por estrangular (*"pinch-off"*) do lado do dreno, como vem representado na Figura 2.4. Sucintamente, aumentando a tensão V_{DS} está a fazer-se com que a junção dreno-substrato esteja mais inversamente polarizada, o que implica que a região de depleção do lado do dreno seja mais profunda e obrigue ao estrangulamento do canal. Nesta situação, o transístor está na região de saturação e a sua característica tensão-corrente é dada pela equação (2.3), onde se mostra que a corrente no dreno é independente da tensão V_{DS}.

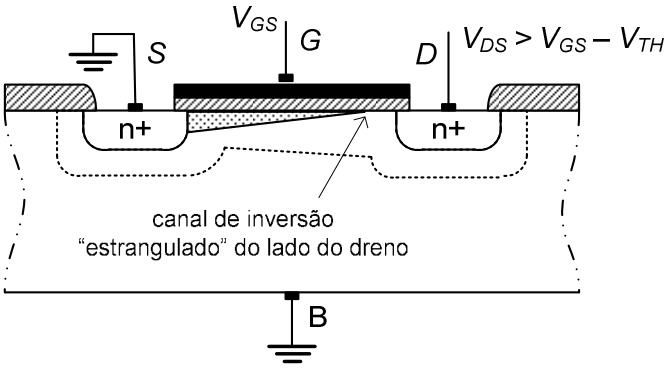

Figura 2.4 – Vista em corte do MOSFET do tipo N na região de saturação.

$$I_{DS} = \frac{k_n'}{2}\frac{W}{L}(V_{GS} - V_{TH})^2 \qquad (2.3)$$

2.1.3 – Efeitos de segunda ordem

O modelo convencional de funcionamento do MOSFET apresentado anteriormente tem vindo, contudo, a sofrer alterações com a evolução da tecnologia. Actualmente, o transístor está sujeito a diversos efeitos secundários, levando a alguns desvios nos modelos de operação do dispositivo. A compreensão dos efeitos de segunda ganha pois maior importância no projecto de circuitos baseados nas tecnologias mais recentes. A avaliação do impacto de alguns destes efeitos, chamados de canal curto, só é conseguida utilizando avançadas ferramentas de simulação, com modelos mais precisos [29].

2.1.3.1 – Corrente de sublimiar

Apesar de se assumir que não existe inversão do canal abaixo da tensão de limiar V_{TH}, a verdade é que, na região de sublimiar, a corrente I_{DS} também aumenta com as tensões V_{DS} e V_{GS}. Significa isto que o transístor já se pode encontrar a conduzir para valores de tensão

inferiores à sua tensão de limiar V_{TH}, ou seja, não começa a conduzir de abruptamente, mas antes de uma forma gradual. Este efeito designa-se por inversão fraca ou condução de sublimiar. A corrente na região de sublimiar depende exponencialmente de V_{GS} e V_{DS}, podendo ser aproximada pela seguinte expressão

$$I_{DS} = I_S \, e^{\frac{V_{GS}}{n k_B T / q}} \left(1 - e^{-\frac{V_{DS}}{k_B T / q}} \right) \tag{2.4}$$

em que I_S (inclinação da característica de corrente) e n são parâmetros empíricos, k_B é a constante de Boltzmann, T a temperatura e q a carga do electrão [29].

Actualmente, devido ao alto nível de integração e ao facto de diversas aplicações requererem um baixo consumo de potência, reduziram-se as tensões de alimentação e, correspondentemente, as tensões de limiar V_{TH}, pelo que os transístores passaram a operar frequentemente junto da região de sublimiar. A influência desta corrente de sublimiar ganha ainda maior importância nos transístores de canal curto, porque a diminuição da V_{TH} aumenta as correntes de fuga nesta região. Nas tecnologias actuais, esta corrente pode assumir valores que vão desde os 0.02 nA/μm, para dispositivos com V_{TH} elevado, até aos 20 nA/μm, para transístores com tensões de limiar mais baixas [29].

2.1.3.2 – Efeito de corpo

Outro dos fenómenos secundários importantes num transístor é o efeito de corpo. A tensão de limiar V_{TH} é função de uma série de parâmetros, como o potencial de Fermi. O efeito de corpo é o termo dado à alteração da tensão V_{TH} quando existe uma diferença de potencial entre a fonte e o substrato V_{SB} [15]. Tendo $V_{SB} > 0$, a região de depleção torna-se maior, implicando uma redução do canal e, consequentemente, da concentração de cargas do canal. Logo, é como se se aumentasse o V_{TH} do transístor, pois este passará a comutar para valores superiores de V_{GS}. Isto é, se a fonte não estiver ao mesmo potencial que o substrato, a tensão V_{SB} tem impacto na tensão de limiar do transístor e o substrato para todos os efeitos age como se fosse uma segunda porta. A influência da tensão V_{SB} na tensão de limiar V_{TH}, sob diferentes condições de ligação do substrato, é então dada por

$$V_{TH} = V_{TH0} + \gamma \left(\sqrt{|\, 2\,\Phi_F \,| + V_{SB}} - \sqrt{|\, 2\,\Phi_F \,|} \right) \tag{2.5}$$

em que V_{TH0} é a tensão de limiar para $V_{SB} = 0$, γ é o factor de efeito de corpo e Φ_F é o potencial de Fermi ($\Phi_F = -0.3$ V para um substrato do tipo P) [29].

2.1.3.3 – Modulação do comprimento do canal

As equações apresentadas anteriormente descrevem o modelo de funcionamento de um MOSFET parecem mostrar que, na saturação, a corrente entre o dreno e a fonte do transístor é constante e independente da tensão aplicada aos seus terminais. Todavia, esta análise não reflecte a verdade, uma vez que não leva em consideração que as alterações na tensão V_{DS} implicam também variações no comprimento do canal [15]. Na região de saturação, o aumento de V_{DS} provoca um aumento da região de depleção junto do dreno e, consequentemente, uma diminuição do comprimento efectivo do canal de inversão. Diz-se, então, que o comprimento do canal é modulado pela tensão V_{DS}.

Para canais longos, a influência da variação no comprimento do canal é pequena. No entanto, à medida que os dispositivos vão sendo reduzidos nas suas dimensões, o comprimento efectivo do canal diminui e essas alterações assumem maior importância. De forma a considerar essa modulação do comprimento do canal, introduz-se uma aproximação para a corrente de dreno do transístor expressa por

$$I_{DS} = \frac{k_n'}{2}\frac{W}{L}(V_{GS} - V_{TH})^2\,(1 + \lambda V_{DS}) \qquad (2.6)$$

onde λ é o factor de modulação do comprimento de canal, inversamente proporcional ao L do dispositivo. O termo linear $1 + \lambda\ V_{DS}$ leva a que a característica tensão-corrente surja ligeiramente alterada. A recta característica da região de saturação a aparecer agora com uma ligeira inclinação, como mostra a Figura 2.5.

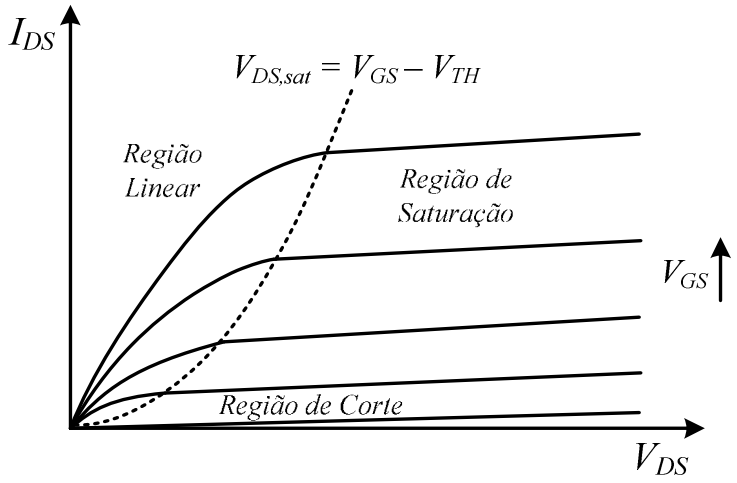

Figura 2.5 – Modulação do comprimento do canal: característica I-V do MOSFET.

2.1.3.4 – Variações da tensão de limiar V_{TH}

Para tecnologias de canal curto, há características físicas intrínsecas aos dispositivos que originam variações significativas nas suas tensões de limiar V_{TH}. Com a redução do comprimento de canal dos transístores, ganham extrema importância as fronteiras efectivas das regiões de depleção em torno do dreno e da fonte. A este nível, um dos fenómenos é conhecido por DIBL (*Drain-Induced Barrier Lowering*) e verifica-se quando há um aumento do potencial entre o dreno e a porta, que vai tornar mais larga a região de depleção em torno do dreno e causa a diminuição da tensão de limiar V_{TH} com o aumento da tensão V_{DS} [29]. Com o aumento da região de depleção em sua volta, o dreno passa a ter controlo sobre a inversão do canal. Como é necessária menos tensão na porta para induzir o canal, origina-se uma redução da tensão de limiar V_{TH} proporcional ao aumento de V_{DS}.

O DIBL é expresso pela equação (2.7) e caracterizado pela variação (em milivolts) provocada na tensão de limiar por cada alteração de um volt na tensão dreno-fonte. Para se ter uma ideia, o DIBL pode implicar variações de 20 a 150 mV na tensão de limiar V_{TH} por cada volt de variação na tensão V_{DS} [29].

$$DIBL = \frac{\Delta V_{TH}}{\Delta V_{DS}} \tag{2.7}$$

Verifica-se ainda que, quanto maior for a tensão aplicada ao dreno, mais para o interior do substrato a região de depleção associada ao dreno aumenta. Num caso extremo, deixa até de existir inversão do canal entre o dreno e a fonte, pois as regiões de depleção aproximam-se e estendem-se até formarem como que uma zona de depleção única. Deste modo, passa a fluir uma corrente no canal independentemente da tensão na porta, uma vez que os portadores "perfuram" o canal de uma região, fenómeno denominado de *punch-through* [29]. O *punch-through* origina um rápido aumento da corrente no transístor e define um limite superior para a tensão V_{DS} que pode ser aplicada ao MOSFET.

2.1.3.5 – Saturação de velocidade

Como já foi referido anteriormente, o comportamento dos transístores nem sempre segue o modelo quadrático de Shichman-Hodges [5]. Actualmente, uma das principais causas – se não a principal – desse desvio nas características tensão-corrente dos transístores MOS é conhecida por efeito de saturação da velocidade de deriva [6]. A velocidade de saturação

pode ser entendida como o limite superior da velocidade dos portadores nos semicondutores e é um dos factores mais sensíveis dos dispositivos de canal curto, para os quais os campos eléctricos são mais intensos [21].

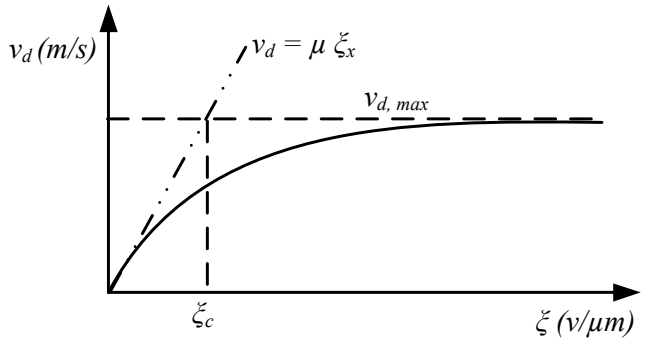

Figura 2.6 – Saturação de velocidade de deriva.

Sabe-se que a velocidade dos portadores relaciona-se com o campo eléctrico de um canal através de um parâmetro designado por mobilidade μ. Numa primeira análise, assume-se que a mobilidade dos portadores é constante e, consequentemente, que a velocidade dos portadores é proporcional ao campo eléctrico, qualquer que seja o valor desse campo. Todavia, para campos eléctricos mais intensos, como nos transístores de canal mais curto, os portadores deixam de seguir um comportamento linear. Assim, a velocidade de deriva dos portadores (v_d) deixa também de depender linearmente do campo eléctrico, para um valor do campo acima de um limiar crítico, ξ_c, como se constata por observação da Figura 2.6 Nestas condições, quando houver um aumento do campo eléctrico, esse efeito não se fará sentir ao nível da velocidade de deriva dos portadores. A velocidade não aumentará, saturando apenas num dado valor constante. A esse fenómeno dá-se o nome de saturação da velocidade de deriva.

Nestas circunstâncias, para um substrato do tipo P, tem-se que a velocidade de saturação dos electrões ronda os 10^5m/s e que o campo eléctrico atinge um valor crítico para um campo de aproximadamente 1.5V/μm. No caso de um substrato do tipo N, é necessário um campo eléctrico superior a 1.5V/μm para que os transístores saturem. Significa isto que os transístores PMOS são menos susceptíveis ao efeito de saturação da velocidade de deriva.

O efeito de saturação da velocidade de deriva tem um impacto considerável no modelo de operação do MOSFET, até porque reduz a quantidade de corrente que o

transístor consegue entregar [10]. Tendo em conta este efeito, a característica tensão-corrente da região linear surge expressa por

$$I_{DS} = k_n \left[(V_{GS} - V_{TH}) V_{DS} - \frac{V_{DS}^{\ 2}}{2} \right] \kappa(V_{DS}) \tag{2.8}$$

com

$$\kappa(V_{DS}) = \frac{1}{1 + \left(\frac{V_{DS}}{\xi_c \, L} \right)} \tag{2.9}$$

em que o quociente V_{DS}/L pode ser tido como o valor do campo eléctrico médio no canal e ξ_c é o valor crítico do campo eléctrico longitudinal ao canal. Relativamente à equação (2.9), o factor $\kappa(V_{DS})$ representa o nível de saturação de velocidade do transístor e a sua análise conduz a uma interpretação simples: para transístores de canal curto, o factor κ é inferior a 1 e o efeito de saturação de velocidade força o transístor a entregar menos corrente [29].

2.1.3.6 – Degradação da mobilidade

Nas tecnologias de canal curto actuais, outro dos efeitos secundários com particular importância é a degradação da mobilidade dos portadores. A mobilidade é descrita como a facilidade com que os portadores se movem num dado material e é definida como a razão entre a velocidade de deriva dos portadores e o campo eléctrico aplicado ao dispositivo. Num transístor, a mobilidade varia com uma série de factores. Varia consoante o tipo de portador de carga – no silício, os electrões têm uma mobilidade superior às lacunas, o que leva a que os dispositivos do tipo N tenham uma capacidade de produção de corrente superior aos do tipo P – e diminui com o aumento da temperatura e com o aumento da concentração de impurezas. Este efeito de segunda ordem caracteriza-se pela redução da mobilidade superficial relativamente à mobilidade dos portadores no interior do substrato.

Nos dispositivos de canal curto, o campo eléctrico no transístor é mais intenso e a mobilidade degrada-se devido à dispersão provocada pelas colisões de portadores com a interface Si-SiO_2, a interface entre o substrato e o óxido da porta [10]. Essas colisões devem-se, sobretudo, ao facto da componente do campo eléctrico transversal (perpendicular ao sentido da corrente) acelerar os electrões, levando-os a colidir com a interface semicondutor-óxido. A mobilidade dos portadores, neste caso, reduz-se. A

mobilidade efectiva apresenta uma dependência com a tensão V_{GS} aplicada ao transístor e vem expressa pela seguinte equação

$$\mu_{eff} = \frac{\mu_0}{1+\theta\,(V_{GS}-V_{TH})} \tag{2.10}$$

onde μ_{eff} é o valor da mobilidade efectiva e θ um valor determinado empiricamente.

2.2 – Capacidades do MOSFET

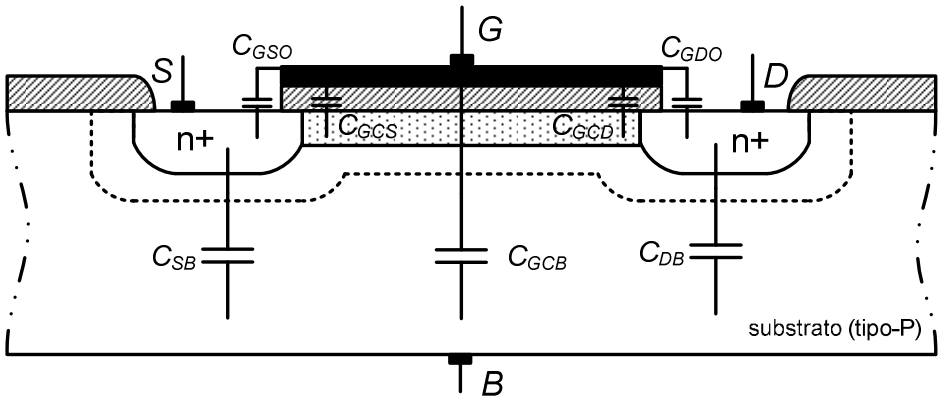

Figura 2.7 – Capacidades do MOSFET: vista em corte do transístor NMOS.

Num transístor, conhecer o conjunto de capacidades intrínsecas dos transístores e das respectivas ligações torna-se essencial para perceber a resposta dinâmica dos mesmos. As capacidades do MOSFET são responsáveis pelo tempo de propagação das portas lógicas, uma das métricas mais importantes de um dado circuito digital. Atente-se na Figura 2.7, relativa à vista em corte do transístor NMOS. Dentro das capacidades dos transístores MOS, identificam-se dois subconjuntos importantes: as capacidades intrínsecas (C_{GCB}, C_{GCS}, C_{GCD}, C_{DB} e C_{SB}) e as capacidades estruturais (de *overlap* da porta).

As capacidades da porta do MOSFET devem-se à separação física entre a porta e o substrato. Deste ponto de vista, podem ser entendidas como as armaduras de um condensador, pelo que estão fortemente relacionadas com a espessura do óxido da porta, t_{ox}, e com a área da região de difusão. A capacidade porta-substrato por unidade de área, C_{ox}, tem profundo impacto nas capacidades intrínsecas dos transístores MOS. A capacidade intrínseca da porta para o canal, C_{GC}, é então definida como a capacidade porta-substrato por unidade de área, C_{ox}, a multiplicar pela área do canal, WL:

$$C_{GC} = C_{ox}WL = \frac{\varepsilon}{t_{ox}}WL \tag{2.11}$$

2.2.1 – Capacidades da porta para o canal

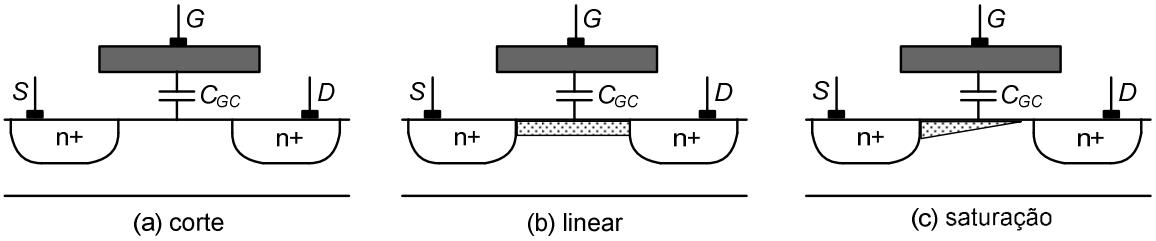

Figura 2.8 – Capacidades da porta para o canal do MOSFET.

A maneira como a capacidade da porta para o canal surge aos terminais do MOSFET depende da inversão do canal, isto é, da região de funcionamento em que o transístor se encontra e da tensão aos terminais do dispositivo. Assim, contribui de forma diferente na imposição dessas capacidades: C_{GCB}, C_{GCS}, C_{GCD} (Figura 2.8). Na região de corte (a), não existe inversão do canal, pelo que é entre a porta e o substrato que surge a capacidade C_{GC}. Para a região linear (b), como já existe um canal entre a fonte e o dreno, a capacidade C_{GCB} é igual a zero e, devido à simetria do transístor, a capacidade C_{GC} está distribuída entre C_{GCS} e C_{GCD}. Por último, na saturação (c), o canal está "estrangulado" do lado do dreno, deixa de existir simetria no dispositivo e a capacidade da porta para o canal, C_{GC}, aparece entre a porta e a fonte [29]. Nesta região, o factor $2/3C_{ox}WL$ é tipicamente a capacidade modelada entre porta e fonte, sendo que os restantes $1/3C_{ox}WL$ são negligenciáveis [11].

2.2.2 – Capacidades de *overlap*

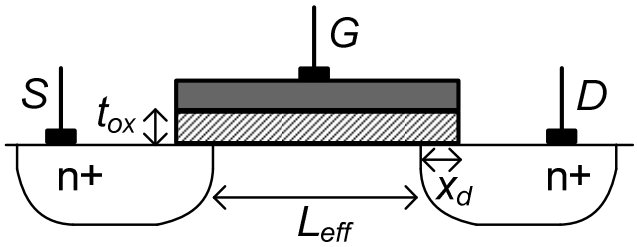

Figura 2.9 – Capacidades de sobreposição da porta com as regiões da fonte e do dreno.

As capacidades de *overlap*, C_{GDO} e C_{GSO}, por sua vez, são as únicas lineares, pois dependem apenas da estrutura física do transístor. As capacidades de *overlap* devem-se, portanto, a uma zona de fronteira em que a porta se sobrepõe, por um lado, à região do dreno e, por outro, à da fonte (Figura 2.9). Devido à indesejável difusão de impurezas, as regiões da fonte e do dreno estendem-se uma porção x_d, denominada por difusão lateral,

para debaixo do óxido da porta, o que implica uma redução do comprimento efectivo do canal L_{eff} de um factor de $2x_d$ [29]. Sendo x_d um parâmetro da tecnologia utilizada, as capacidades vêm dadas pela equação (2.12). Como x_d representa um valor fixo para determinado processo, pode ser combinado com C_{ox}, resultando na capacidade por unidade de largura C_o.

$$C_{GSO} = C_{GDO} = C_{ox}\, x_d\, W = C_o\, W \tag{2.12}$$

Para estimar convenientemente as capacidades dos transístores, apresenta-se na Tabela 1 um sumário da distribuição da capacidade da porta, C_G, para as diferentes regiões de funcionamento do MOSFET.

Região	C_{GCB}	C_{GCS}	C_{GCD}	C_{GC}	C_G
Corte	$C_{ox}WL$	0	0	$C_{ox}WL$	$C_{ox}WL + 2C_oW$
Linear	0	$C_{ox}WL / 2$	$C_{ox}WL / 2$	$C_{ox}WL$	$C_{ox}WL + 2C_oW$
Saturação	0	$(2/3)C_{ox}WL$	0	$(2/3)C_{ox}WL$	$(2/3)C_{ox}WL + 2C_oW$

Tabela 1.1 – Capacidades do MOSFET dependentes do canal de inversão

2.2.3 – Capacidades das junções PN

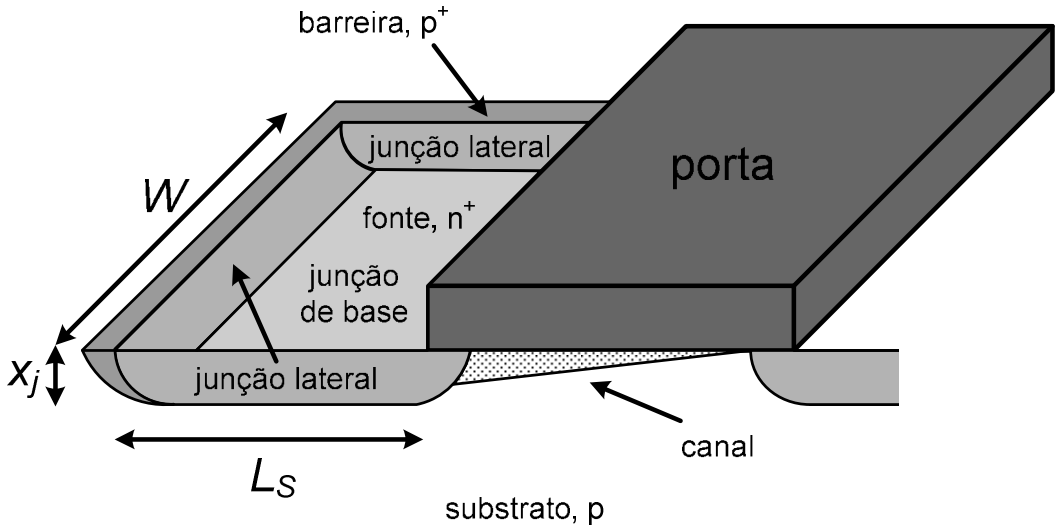

Figura 2.10 – Vista em detalhe da junção PN da fonte do MOSFET.

Outra das componentes das capacidades intrínsecas do MOSFET (Figura 2.10) está relacionada com as regiões de depleção que envolvem a fonte e o dreno, regiões essas onde se formam com o substrato junções PN inversamente polarizadas [26]. Estas capacidades

das junções PN (fonte-substrato, C_{SB}, e dreno-substrato, C_{DB}) formam-se devido às diferentes dopagens do substrato, p, da fonte, n^+, e da barreira lateral, p^+, e incluem o efeito de duas junções: a junção da base e a junção lateral [29]. A expressão para a capacidade total das junções PN, também designada por capacidade de difusão, é a seguinte

$$C_{diff} = C_j L_s W + C_{jsw}(W + 2L_s) \tag{2.13}$$

em que L_s é o comprimento da junção lateral, C_j a capacidade de área da junção da base e C_{jsw} a capacidade por unidade de perímetro da junção lateral da região de depleção do MOSFET. De referir que as capacidades C_j e C_{jsw} dependem exponencialmente das tensões aplicadas à junção, uma vez que se tratam de capacidades de depleção. Uma expressão genérica para a capacidade da junção C_j é dada pela equação (2.14)

$$C_j = C_{j0}\left(1 - \frac{V_B}{\Phi_0}\right)^{-m_B} \tag{2.14}$$

onde C_{j0} é a capacidade da junção sob condições de polarização nula, V_B a tensão aplicada à junção, Φ_0 o potencial da junção e m_B o coeficiente que classifica o tipo de junção quanto à dependência com a tensão aplicada [20, 29].

2.2.4 – Modelo de capacidades do MOSFET

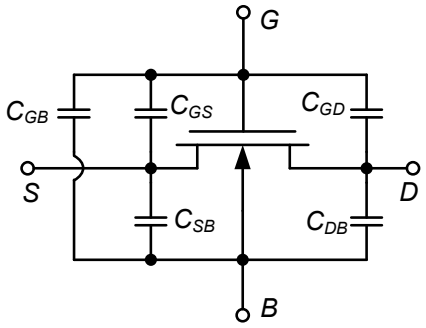

Figura 2.11 – Modelo de capacidades do MOSFET.

Juntando todas as contribuições apresentadas anteriormente, tem-se o modelo de capacidades do transístor do tipo N apresentado na Figura 2.11. A capacidade porta-fonte C_{GS} é a soma das capacidades C_{GCS} e C_{GSO} e, de forma análoga, a capacidade porta-dreno C_{GD} é a soma das contribuições C_{GCD} e C_{GDO}. As capacidades fonte-substrato C_{SB} e dreno-substrato C_{DB} são as capacidades dependentes da região de depleção das junções PN.

2.2.5 – Comportamento das ligações

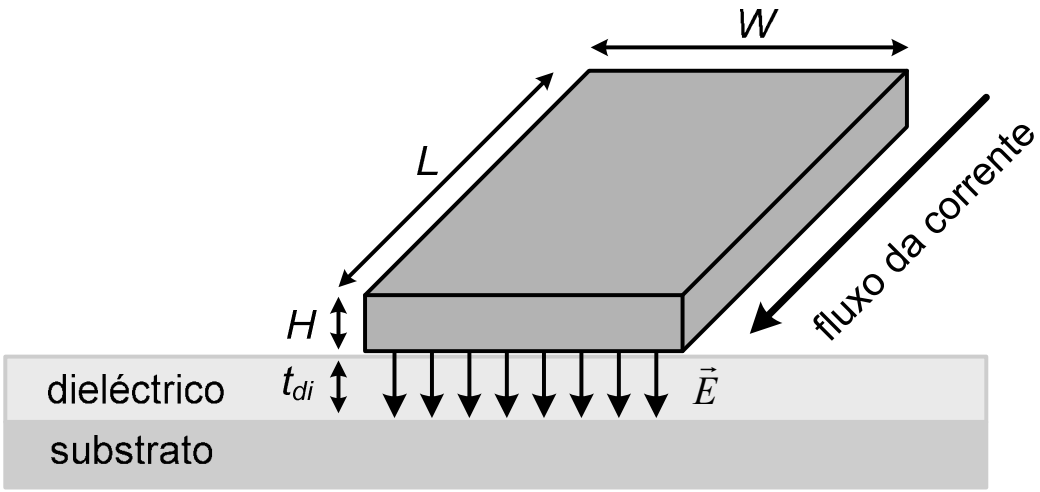

Figura 2.12 – Estrutura tridimensional de uma ligação.

Num circuito integrado, as ligações entre componentes exibem capacidades parasitas. Cada ligação é uma estrutura tridimensional (Figura 2.12) de metal e/ou polisilício com variações significativas na sua forma, espessura ou distância para o substrato [11]. Para além disso, cada ligação está rodeada de outras linhas, no mesmo nível ou em níveis diferentes (Figura 2.13). Assim, à medida que a tecnologia adopta dispositivos de dimensões cada vez mais reduzidas, o efeito destas ligações torna-se um problema cada vez mais significativo. Prevê-se que, quando a largura das ligações, *W*, se tornar mais pequena que 1.75 vezes a espessura destas, *H*, a capacidade entre fios começará a ser dominante [29]. Ganha, portanto, maior importância a correcta estimação das capacidades parasitas desses fios.

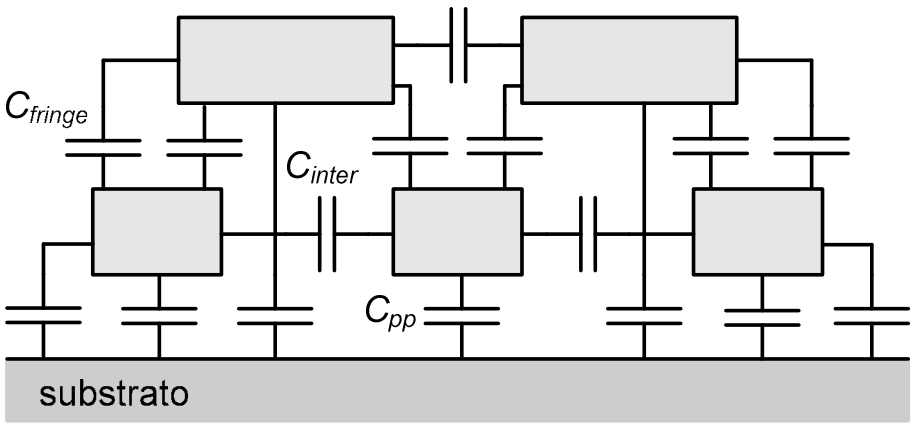

Figura 2.13 – Vista em detalhe das camadas de ligações entre dispositivos.

Numa primeira análise considera-se que as ligações formam condensadores de placas paralelas entre si, capacidade designada por C_{pp} e dada por

$$C_{pp} = \frac{\varepsilon_{di}}{t_{di}} WL \tag{2.15}$$

em que ε_{di} representa a permitividade do dieléctrico e t_{di} a espessura do dieléctrico.

No entanto, a evolução da escala de integração (e consequente redução da espessura do óxido e do espaçamento entre os metais) levou a que seja necessário ponderar, para além dessa capacidade, o efeito da capacidade lateral (C_{fringe}) e, principalmente, das capacidades entre fios (C_{inter}). Assim, a influência das capacidades parasitas das ligações pode ser aproximada de forma mais exacta com base na seguinte equação [26]

$$C_{wire} = \varepsilon_{di} \left[\frac{W}{t_{di}} - \frac{H}{2t_{di}} + \frac{2\pi}{\ln\left(1 + \frac{2t_{di}}{H}\left\{1 + \sqrt{1 + \frac{H}{t_{di}}}\right\}\right)} \right] \times L \tag{2.16}$$

onde W, L e H são, respectivamente, a largura, o comprimento e a espessura das ligações. Segundo outros autores [19, 29], pode utilizar-se um outro modelo simplificado para aproximar esta capacidade de ligação, conseguindo-se com $w = W - H/2$ uma aproximação prática bastante aceitável para o condensador de placas paralelas C_{pp}. Essa aproximação é expressa pela equação (2.17).

$$C_{wire} = C_{pp} + C_{fringe} = \frac{w\,\varepsilon_{di}}{t_{di}} + \frac{2\pi\,\varepsilon_{di}}{\log\left(\frac{t_{di}}{H}\right)} \tag{2.17}$$

Para fazer estimativas "à mão" dos valores destas capacidades parasitas das ligações, os fabricantes facultam uma série de tabelas com um conjunto de valores típicos para as diversas camadas de ligação de determinado processo tecnológico [29]. Nessas tabelas é, por norma, feita a distinção entre as capacidades de área (expressas em $aF/\mu m^2$) e as capacidades ditas laterais (em $aF/\mu m$). Os valores tabelados, que vêm expressos para uma distância mínima entre as ligações, são obtidos à custa de programas e ferramentas avançadas de simulação.

2.3 – Teoria de *scaling*

Nas últimas décadas, assistiu-se a um enorme desenvolvimento da indústria dos semicondutores, motivada pela evolução dos processos tecnológicos e pela evolução ao nível do fabrico dos dispositivos [38]. A densidade de transístores por circuito integrado aumentou para patamares inimagináveis e levou a que esses transístores fossem sendo criados com dimensões tão pequenas quanto o possível. Essa redução sistemática das dimensões físicas do transístor entende-se por *scaling* e vem ilustrada na Figura 2.14.

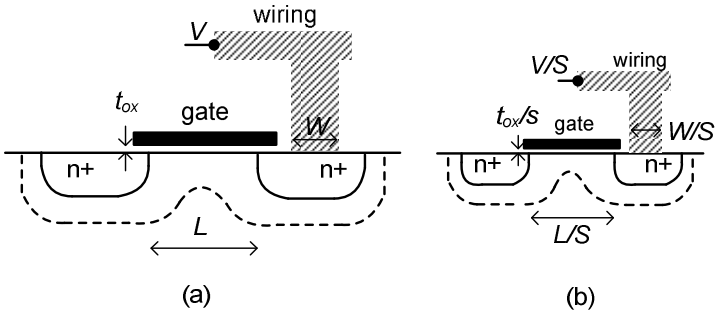

Figura 2.14 – Princípios do *scaling*: dispositivo original (a) e reduzido (b).

As primeiras teorias de *scaling* [7, 8] surgiram na década de 70, motivadas essencialmente por factores económicos e de desempenho eléctrico. Originaram um grande desenvolvimento das tecnologias MOS e foram um meio de sustentar a Lei de Moore, possibilitando o crescimento das aplicações VLSI. Além disso, o *scaling*, por si só, é especialmente relevante no campo da predição, servindo como guia de orientação para o projecto de novas gerações tecnológicas a partir de uma dada geração em voga. Através da denominada análise de *scaling*, pode prever-se de que forma a tendência de desenvolvimento da tecnologia irá afectar os principais requisitos de um circuito integrado e quais os desafios que essa redução de dimensões irá trazer no que respeita a tentar manter os circuitos realizáveis na prática. Na secção que se segue, são abordados os três diferentes cenários de *scaling*, bem como a sua influência no modelo de funcionamento do MOSFET e nas principais métricas dos circuitos digitais.

2.3.1 – *Scaling* de campo eléctrico constante

Baseado num modelo formulado por Dennard no início da década de 70 [7, 8], o *scaling* de campo eléctrico constante (ou *scaling* completo) foi a primeira teoria proposta. Segundo este modelo ideal, mantendo constante o campo eléctrico no transístor, reduzem-se, tal como mostra a Figura 2.14, todas as dimensões (W, L e t_{ox}) e tensões (V_{DD} e V_{TH}) do

transístor de um factor *S*. Como o campo eléctrico se mantém constante, este processo de *scaling* apresenta a vantagem de evitar o aparecimento de efeitos secundários ao modelo do transístor, como a degradação da mobilidade ou a saturação de velocidade. Isto assumindo que a alteração nas dopagens N_D e N_A do transístor não afectam a mobilidade dos portadores de carga [18, 19].

Este modelo de *scaling* é vantajoso fundamentalmente por dois motivos: o desempenho do transístor melhora, uma vez que o seu tempo de resposta diminui de um factor *S*, enquanto a sua potência dissipada diminuiu de forma quadrática, de um factor de S^2. A mais atractiva das vantagens é mesmo a possibilidade de se ter um maior número de circuitos acomodados numa dada área de um *chip*, sem implicações ao nível da densidade de potência deste. No entanto, este processo não trouxe só benefícios. A redução constante das tensões leva a que, no que respeita à tensão de limiar, se esteja a aumentar a condução da região de sublimiar e as correntes de fugas. Além disso, apesar de se manter um conceito válido, o *scaling* de campo eléctrico constante deixou de ser utilizado devido ao facto de reduzir as tensões em proporção com as reduções das dimensões. Na altura, pretendia-se estabelecer valores padrão para a interface com o *chip*, o que nunca se conseguiria continuando a adoptar este método.

2.3.2 – *Scaling* de tensões fixas

Dadas as dificuldades práticas levantadas pelo método anterior, foram colocadas em prática outras formas de abordagem ao *scaling*, sendo introduzidos outros métodos. O denominado *scaling* de tensões fixas foi o que se seguiu, tendo sido um modelo utilizado até ao início da década de 90. Neste caso, as dimensões (*W*, *L* e X_j) são reduzidas de um factor *S*, mas a compatibilidade das tensões é mantida, preservando-se a tensão de alimentação V_{DD} num determinado patamar e outras tensões durante o processo [15]. Em 1990, o valor padrão para o nível de sinal de alimentação era 5 V, mas entretanto foram sendo introduzidos novos padrões e hoje em dia, por exemplo, na tecnologia CMOS de 90 nm, a tensão de alimentação é de 1.0 V. Por essas razões, prefere-se o modelo de tensões fixas ao *scaling* de campo eléctrico constante.

No *scaling* de tensões fixas, as tensões permanecem inalteráveis, mas todas as dimensões do MOSFET são reduzidas de um factor *S*. Todavia, é boa prática que a

espessura do óxido t_{ox}, não seja reduzida do mesmo factor, pois provocaria um aumento significativo no campo eléctrico do transístor, causando degradação da mobilidade [18]. De forma a atenuar esse problema, a t_{ox} é de certa forma "menos reduzida". Mesmo assim, o problema do aumento do campo eléctrico do dispositivo no modelo de tensões fixas tem sempre que ser tido como uma desvantagem do processo, pois, contrariamente ao método de campo eléctrico constante, tem implicações relativamente aos efeitos de segunda ordem do MOSFET.

2.3.3 – *Scaling* geral

Apesar de outros tipos de *scaling* poderem ser aplicados, como o de tensões quase constantes ou o *scaling* lateral (conhecido também por apenas o comprimento da porta ser reduzido), vários estudos levaram à construção de um modelo mais realista [29]. Observando a evolução dos processos tecnológicos, nota-se que o *scaling* da tensão de alimentação dos circuitos não tem acompanhado o da tecnologia. Por exemplo, para a recente tecnologia de 90 nm, a tensão de alimentação é de 1.0 V, como já foi referido no ponto anterior, enquanto para 0.8 μm a tensão de alimentação era de 5 V. Motivado pela impossibilidade de derrubar alguns dos parâmetros intrínsecos aos dispositivos, não escaláveis, como o potencial de Fermi Φ_F, surgiu então um modelo mais realista designado por *scaling* geral, no qual as dimensões e as tensões são reduzidas de factores independentes.

Neste processo de *scaling* mais geral, as dimensões do transístor são reduzidas de um factor S, ao passo que as tensões são reduzidas de um factor U. Caso $U = 1$, isto é, caso a tensão seja mantida constante, tem-se o método de tensões fixas. Depreende-se então que o método geral é o melhor de todos em termos de performance, pois, relativamente aos outros dois métodos, conjuga uma redução na potência dissipada com uma melhoria ao nível da densidade de potência, característica de extrema importância em circuitos VLSI.

A Tabela 1.2. resume os três diferentes tipos de *scaling* apresentados anteriormente, mostrando o efeito desses modelos nos parâmetros do dispositivo e nas métricas mais importantes dos circuitos digitais. Numa análise de *scaling*, é dada especial relevância às dimensões e tensões dos dispositivos, bem como à densidade de potência e à potência dissipada pelos circuitos [29].

Parâmetro	Expressão	Modelo de *Scaling*		
		Completo	**Tensões Fixas**	**Geral**
W, L, t_{ox}		$1/S$	$1/S$	$1/S$
V_{DD}, V_{TH}		$1/S$	1	$1/U$
I_{DS}		$1/S$	1	$1/U$
N_D, N_A	V/W_{depl}^2	S	S^2	S^2/U
k'_n, k'_p		$1/S^2$	$1/S^2$	S
k_n, k_p		S	S	S
C_{ox}	$1/t_{ox}$	S	S	S
C_G	$C_{ox}WL$	$1/S$	$1/S$	$1/S$
t_p	$R_{on}C_G$	$1/S$	$1/S$	$1/S$
Área	WL	$1/S^2$	$1/S^2$	$1/S^2$
P_{av}	$I_{DS}V$	$1/S^2$	1	$1/U^2$
Densidade Potência	$P_{av}/(WL)$	1	S^2	S^2/U^2

Tabela 1.2 – Influência do *scaling* nos diferentes parâmetros do MOSFET.

2.3.4 – Impacto do *scaling* na física dos dispositivos MOS

O progressivo *scaling* da tecnologia tem impacto a toda a linha sobre a física dos dispositivos. Para tecnologias de dimensões mais reduzidas, os chamados efeitos de canal curto ganham relevância e afectam a própria característica corrente-tensão do MOSFET. A Figura 2.15 ilustra isso mesmo, comparando o comportamento I-V em transístores com canais longos e com canais curtos.

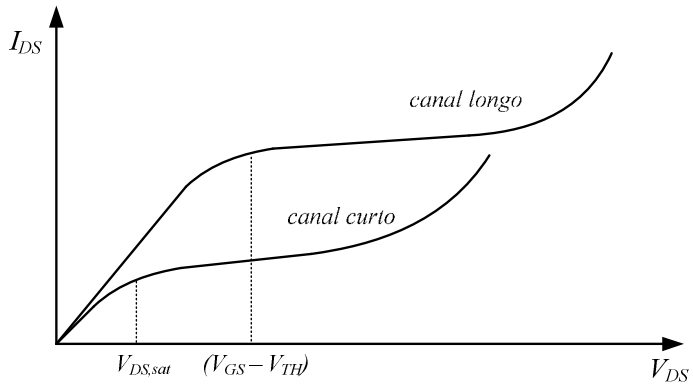

Figura 2.15 – Transístores de canais longo e curto: comparação entre características I-V.

Através desta comparação, a principal conclusão que se tira reside no facto dos transístores de canal curto apresentarem uma região de saturação maior que os dispositivos

de canal longo [29]. Isto sucede porque a tensão de saturação $V_{DS,sat}$ é inferior ao patamar $V_{GS} - V_{TH}$, logo o dispositivo de canal curto opera mais na saturação. Diz-se então que, para os transístores actuais, a saturação de velocidade antecipa a saturação para valores inferiores de V_{DS} e leva a que o transístor sature sem ser por *pinch-off*. Assim, reduz-se a quantidade de corrente que pode ser entregue e, na região de saturação, passa a notar-se uma dependência linear da corrente I_{DS} com a tensão V_{GS}. Nestas condições, como mostra a Figura 2.15, o transístor apresenta na sua característica problemas de *punch-through* para valores menores da tensão V_{DS} [10, 19].

Como um dispositivo é designado de canal curto se o comprimento do canal é da mesma ordem de magnitude que as regiões de depleção do dreno e da fonte [19], a constante redução das dimensões leva a que fenómenos secundários como o DIBL ou o *punch-through* se tornem cada vez mais prováveis. Isto para além do impacto ao nível das tensões de operação do transístor. A tensão de alimentação vem sendo continuamente reduzida e isso tem originado uma redução agressiva na tensão de limiar V_{TH}, o que implica um aumento da condução na região de sublimiar e um correspondente aumento das correntes de fuga. Para lidar com estes e outros efeitos nefastos do *scaling*, foram feitos esforços no sentido de adoptar novas estruturas físicas como alternativas ao CMOS convencional e a introduzir novos materiais, tais como os dieléctricos "high-k" [19, 35]. É neste contexto que surge a optimização ao nível dos dispositivos, que é, seguramente, um dos maiores desafios que a indústria dos semicondutores terá de vencer num futuro próximo [34].

Consequentemente, para expandir as fronteiras iminentes da constante miniaturização dos dispositivos, apareceu a tecnologia SOI (*Silicon-On-Insulator*) como alternativa às estruturas tradicionais [27, 29]. Uma estrutura SOI é concebida através da deposição de uma fina camada de óxido no substrato do dispositivo físico, o que possibilita um isolamento vertical do mesmo, pois esse BOX (*Buried Oxide*) serve para isolar dielectricamente o transístor. É desse isolamento que provém a grande vantagem das estruturas SOI: providenciam um melhor controlo sobre fenómenos como as correntes de fuga, optimizando as características de operação do dispositivo [27].

Na última década, foi desenvolvida uma grande variedade de estruturas bastante diferentes das mais familiares. De entre as estruturas SOI estado-da-arte que combatem de maneira mais eficaz os problemas da constante miniaturização, destacam-se duas, que se

apresentam na Figura 2.16: o transístor GAA (*Gate All-Around*) [31] e o FinFET [27], duas das estruturas que se crê que serão as melhores soluções para a implementação dos processos tecnológicos futuros [27]. Outros estudos [34] apresentam alternativas como o DGSOI (*Double-Gate SOI*), um transístor com uma segunda porta, ou o DTMOS (*Dynamic Threshold MOS*), uma estrutura em que a porta está ligada ao substrato do SOI, o que possibilita que o transístor tenha uma tensão de limiar baixa quando está no estado activo e uma tensão de limiar elevada quando está ao corte – redução efectiva da corrente de fugas.

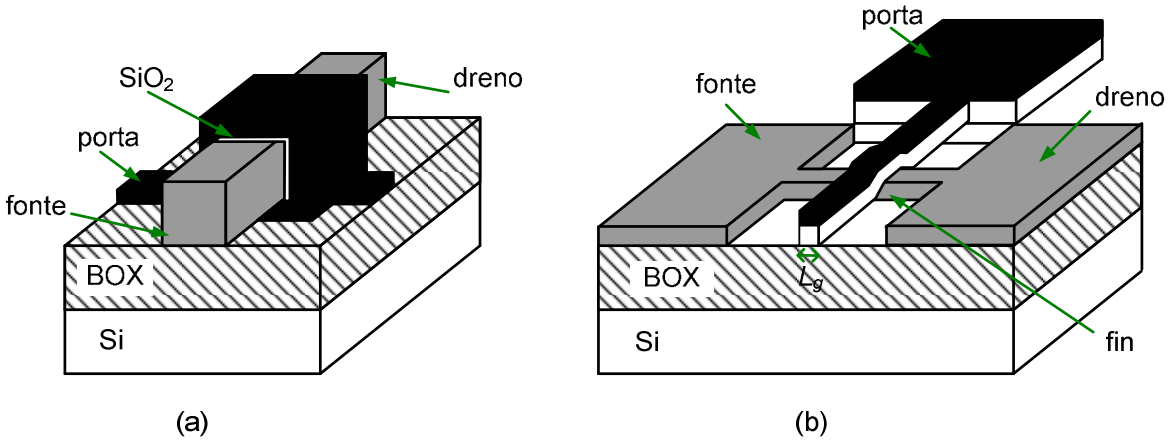

Figura 2.16 – Estruturas SOI: (a) GAA MOSFET e (b) FinFET.

Nos últimos anos, os investigadores encontraram o nanotubo de carbono, um material que entendem ser mais forte que o aço, mais leve que o alumínio e melhor condutor que o cobre [32]. Pode ser utilizado quer na própria estrutura do transístor, quer como ligações entre transístores, o que faz do nanotubo de carbono o material apropriado para os circuitos integrados da nova geração. A aposta dos investigadores é que o transístor de nanotubos de carbono [30, 32] e acreditam que essa estrutura a médio prazo irá substituir o MOSFET convencional. Neste tipo de estrutura, é colocado um nanotubo de carbono entre o dreno e fonte. Esse minúsculo nanotubo, cujas paredes têm apenas um átomo de carbono de espessura, desempenha o papel de canal do transístor e está isolado do resto do dispositivo, possibilitando um desempenho bastante superior ao transístor tradicional de silício (crê-se que 10 a 100 vezes superior) [30].

2.4 – Sumário

Nas disposições finais deste capítulo convém realçar o carácter introdutório do mesmo. Como o transístor está na génese da construção de qualquer circuito digital, neste capítulo foram apresentados os princípios básicos do modelo geral do MOSFET, as suas características de funcionamento e discutidos os efeitos secundários mais importantes, focando principalmente os chamados efeitos de canal curto como o DIBL e a saturação de velocidade de deriva. De seguida, aborda-se o tema das capacidades do MOSFET. Apresentam-se as capacidades intrínsecas dos transístores e identificam-se os subconjuntos que têm maior influência na resposta dinâmica de um circuito.

No contexto da física do dispositivo, é igualmente apresentada a teoria de *scaling* e os seus diferentes conceitos. Na perspectiva da tecnologia, é comentada a necessidade de adoptar novas estruturas para fazer face aos desafios criados pela constante miniaturização dos transístores. Ganhou relevância a chamada optimização ao nível do dispositivo. São apresentadas algumas das estruturas alternativas aos dispositivos CMOS convencionais, com destaque para o FinFET e para os transístores de nanotubos de carbono.

Capítulo 3

Desenho de Circuitos CMOS Estáticos

Neste capítulo, são alvo de discussão as diversas técnicas de desenho utilizadas no projecto de circuitos CMOS, desde o dimensionamento do inversor estático até ao de outras portas, com múltiplas entradas. Proposta pela primeira vez no início da década de 60 [3], a tecnologia CMOS desempenhou um papel fulcral na evolução da microelectrónica. Hoje em dia, é claramente a tecnologia dominante no desenho de circuitos integrados.

Porém, as principais técnicas de desenho de circuitos foram formuladas numa altura em que os transístores apresentavam características de operação bem definidas, em contraste com as tecnologias mais recentes. Assim, torna-se relevante estudar o comportamento destas mesmas técnicas com o avanço da tecnologia e a sua aplicação ao desenho de circuitos que utilizem transístores de canal curto como os actuais. Ao longo deste capítulo, são discutidos os modelos analíticos formulados para ultrapassar as limitações dos modelos convencionais do desenho em CMOS. Os estudos de Hedenstierna e Jeppson [9] postularam as primeiras derivações ao modelo tradicional, mas são igualmente alvo de análise outros modelos.

3.1 – O inversor CMOS

Dentro da tecnologia CMOS, o inversor é tido como o ponto de partida para o desenho dos mais diversos circuitos digitais. Conhecidas as suas propriedades e analisadas as suas

características de operação, facilita-se a análise de circuitos mais complexos, que podem ser estudados mediante uma extensão da análise feita para o inversor. A Figura 3.1 mostra a configuração típica de um inversor CMOS estático, composto por um transístor PMOS e um NMOS ligados em série e com uma capacidade de carga C_L na saída (que representa a carga concentrada na saída do circuito). Quanto ao modo de operação, o inversor pode ser entendido através do seguinte modelo simplificado: quando a entrada do circuito se encontra a um nível baixo, o PMOS faz com que a saída seja colocada a V_{DD}, o nível lógico "1"; quando a entrada está a um nível alto, é o NMOS que está activo e que coloca a saída no nível lógico "0".

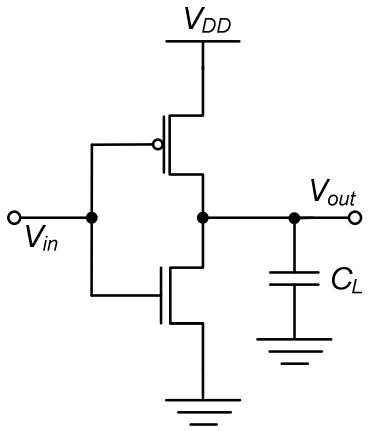

Figura 3.1 – O inversor CMOS.

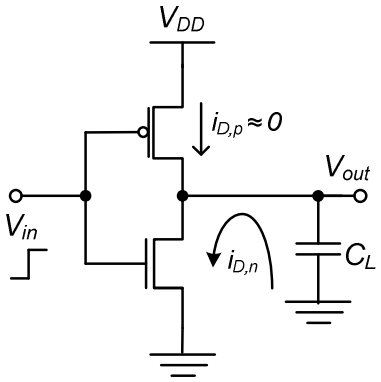

Figura 3.2 – Inversor durante a descarga.

Através do modelo de análise das correntes nos transístores e na capacidade de carga (Figura 3.2), torna-se mais fácil o estudo das características dos circuitos digitais no domínio do tempo. A resposta transitória do inversor CMOS, por sua vez, é dominada pelos tempos de propagação, que definem o atraso que determinado sinal sofre ao passar por uma porta lógica. Os tempos de propagação, numa primeira aproximação, estão relacionados com o tempo que demora a carga e a descarga da capacidade de carga C_L, podendo ser determinados através da solução da equação de estado do nó de saída. A equação diferencial da corrente i_c através do condensador C_L, no domínio do tempo, é dada por

$$i_c(t) = C_L \cdot \frac{dV_{out}}{dt} = i_{D,p} - i_{D,n} \tag{3.1}$$

sendo i_{Dp} e i_{Dn} as correntes de dreno do PMOS e do NMOS, respectivamente [12, 13, 14].

Numa primeira análise, considera-se um sinal de estímulo na entrada a transitar abruptamente entre 0 e V_{DD}. Para esta transição ascendente da entrada, o transístor NMOS encontra-se no seu estado activo e começa a descarregar a capacidade de carga, enquanto o

PMOS está ao corte ($i_{D,p} \approx 0$). A equação de corrente que descreve o processo de descarga é então a seguinte:

$$C_L \cdot \frac{dV_{out}}{dt} = -i_{D,n} \tag{3.2}$$

O tempo de propagação t_{pHL} define o tempo de resposta da porta para uma transição descendente na saída, que corresponde a uma transição ascendente na entrada. Nestas circunstâncias, o NMOS opera ora na saturação ora região linear. Por integração da corrente $i_{D,n}$ nas duas regiões de funcionamento, tem-se que o tempo de propagação de descida t_{pHL} é dado pela soma das duas contribuições

$$t_{pHL} = \frac{C_L}{k_n (V_{DD} - V_{THn})}\left[\frac{2 V_{THn}}{V_{DD} - V_{THn}} + \ln\left(\frac{4 (V_{DD} - V_{THn})}{V_{DD}} - 1\right)\right] \tag{3.3}$$

sendo C_L a capacidade concentrada na saída do inversor, V_{DD} a tensão de alimentação do circuito, k_n a transcondutância do transístor NMOS e V_{THn} a sua tensão de limiar [19, 36].

De forma análoga, considerando a entrada a transitar entre V_{DD} e 0, tem-se que para esta transição descendente na entrada o transístor PMOS está activo e inicia o processo de carga da capacidade C_L. O NMOS está cortado ($i_{D,n} \approx 0$) e não entra na análise, como mostra o modelo da Figura 3.3. A equação de corrente que descreve o evento de carga é, portanto, a seguinte:

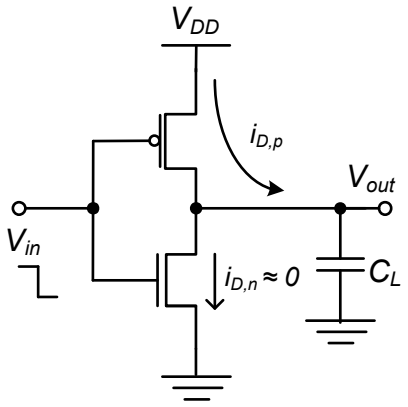

Figura 3.3 – Inversor durante a carga.

$$i_{D,p} = C_L \cdot \frac{dV_{out}}{dt} \tag{3.4}$$

Seguindo um método semelhante ao anterior, nestas condições determina-se o tempo de propagação t_{pLH}, que se refere ao tempo entre a transição negativa na entrada e correspondente transição ascendente na saída. A integração da equação diferencial (3.4) deve ser resolvida para cada uma das regiões de funcionamento do inversor e resulta em

$$t_{pLH} = \frac{C_L}{k_p (V_{DD} - |V_{THp}|)}\left[\frac{2 |V_{THp}|}{V_{DD} - |V_{THp}|} + \ln\left(\frac{4 (V_{DD} - |V_{THp}|)}{V_{DD}} - 1\right)\right] \tag{3.5}$$

em que k_p é a transcondutância do transístor PMOS e V_{THp} a sua tensão de limiar [19].

3.2 – Técnicas tradicionais de desenho

No inversor estático, a técnica mais convencional de desenho em CMOS consiste na manipulação das dimensões relativas de cada um dos transístores NMOS e PMOS que o constituem. O princípio base centra-se no equilíbrio entre as resistências equivalentes dos transístores de tipo P e de tipo N [15]. É mesmo a técnica mais efectiva que se pode utilizar, pois as dimensões *W* e *L* são, para um dado processo CMOS, os únicos parâmetros cujo controlo está ao alcance do desenhador. Dimensionam-se os transístores do inversor de modo a fazer corresponder as resistências dos PMOS às do NMOS, sendo que esta aproximação permite obter uma característica de transferência com elevada simetria e, simultaneamente, equilibrar os tempos de propagação do circuito [11].

Nos circuitos CMOS, os atrasos de propagação caracterizam-se pelo pior caso. As técnicas de desenho consistem invariavelmente nas aproximações necessárias para minimizar a perda de eficiência provocada nesse pior caso. Numa perspectiva ideal, apenas um transístor do inversor conduz durante cada transição na entrada, enquanto o outro transístor está cortado. Através dos modelos simplificados da Figura 3.4, a condição que possibilita a obtenção de tempos de propagação de subida t_{pLH} e de descida t_{pHL} iguais é facilmente inteligível. Através das expressões (3.3) e (3.5), tem-se que, para $| V_{THp} | = V_{THn}$, o equilíbrio entre tempos de propagação reduz-se à correspondência entre as transcondutâncias do NMOS e do PMOS. Esta correspondência implica equilibrar as resistências equivalentes R_{eqN} e R_{eqP}, que são inversamente proporcionais à razão *W/L* dos dispositivos [11], e expressas pelas aproximações seguintes:

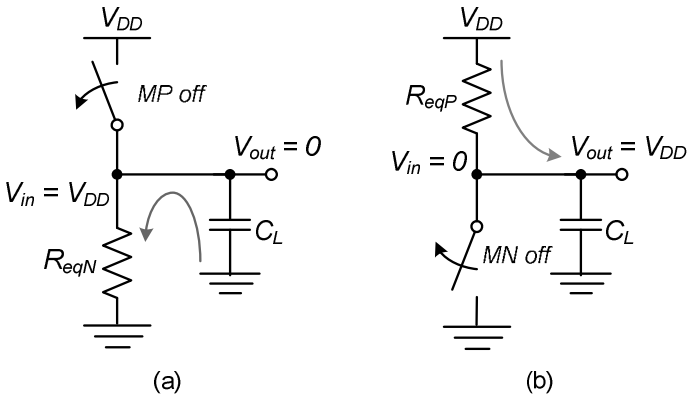

Figura 3.4 – Modelo simplificado do inversor para entrada alta (a) e baixa (b).

$$R_{eqN} \propto \frac{1}{k'_n \left(\frac{W}{L}\right)_N} \tag{3.6}$$

$$R_{eqP} \propto \frac{1}{k'_p \left(\frac{W}{L}\right)_P} \tag{3.7}$$

O parâmetro de transcondutância *k'* do MOSFET é função da mobilidade μ dos portadores de carga

$$k' = \frac{\mu C_{ox}}{2} \tag{3.8}$$

em que C_{ox} é a capacidade por unidade de área do óxido da porta do transístor. De acordo com as características intrínsecas dos transístores, o parâmetro k_n' do NMOS é, por norma, duas a quatro vezes superior ao parâmetro k_p', devido às diferenças na mobilidade dos portadores dominantes dos dois dispositivos. Deste modo, para compensar o facto do valor típico da mobilidade das lacunas μ_p ser inferior à dos electrões μ_n, o PMOS deva ser significativamente mais largo que o NMOS, para a razão das transcondutâncias k_n/k_p ser aproximadamente igual a 1. Assim, para uma dada tecnologia de comprimento L fixo, colocando de parte efeitos secundários como a degradação da mobilidade [29, 36], a razão μ_n/μ_p é constante e pode ser compensada através do ajuste dos parâmetros de desenho W_p e W_n. É nestas considerações que assenta a mais convencional técnica de desenho de circuitos digitais.

$$R_{eqP} = R_{eqN} \quad \Rightarrow \quad \beta = \frac{k_n'}{k_p'} = \frac{\left(\frac{W}{L}\right)_P}{\left(\frac{W}{L}\right)_N} \quad \Leftrightarrow \quad \beta = \frac{\mu_n}{\mu_p} = \frac{W_p}{W_n} \tag{3.9}$$

Da correspondência descrita em (3.9), tem-se portanto que os transístores PMOS devem ser β vezes mais largos que os NMOS de modo a compensar as diferenças na mobilidade. Desta maneira, estão também a equilibrar-se os tempos de propagação.

3.2.1 – Influência das diferenças entre tensões de limiar

No entanto, na tecnologia CMOS, os transístores NMOS e PMOS não são concebidos com tensões de limiar iguais. Nas tecnologias de canal curto actuais, é mais significativa a diferença entre as tensões V_{THn} e V_{THp}, pelo que a técnica de desenho tradicional enunciada anteriormente deixa de ser tão efectiva como era outrora, quando foi formulada. Olhando para as equações (3.3) e (3.5), como as tensões de limiar não são iguais, o cálculo do β que

conduz a um desenho mais próximo do desenho optimizado tem que ser ajustado com a introdução na equação do factor $A(V_{THn},V_{THp})$ [36]. Consequentemente,

$$\beta = \frac{W_p}{W_n} = A(V_{THn},V_{THp}) \cdot \frac{\mu_n}{\mu_p} \tag{3.10}$$

sendo o factor $A(V_{THn},V_{THp})$ função das tensões V_{DD}, V_{THn} e V_{THp}, dado por:

$$A(V_{THn},V_{THp}) = \frac{(V_{DD}-V_{THn})}{(V_{DD}-|V_{THp}|)} \cdot \frac{\left[\frac{2|V_{THp}|}{V_{DD}-|V_{THp}|} + \ln\left(\frac{4(V_{DD}-|V_{THp}|)}{V_{DD}} - 1\right)\right]}{\left[\frac{2V_{THn}}{V_{DD}-V_{THn}} + \ln\left(\frac{4(V_{DD}-V_{THn})}{V_{DD}} - 1\right)\right]} \tag{3.11}$$

Deste modo, ao invés de estar apenas dependente da mobilidade dos portadores, a razão W_p/W_n óptima passa a ser igualmente função das tensões de limiar V_{THp} e V_{THn} dos dois transístores que compõem o inversor e da tensão de alimentação V_{DD}.

3.2.2 – Influência da capacidade de carga e da transição da entrada

As técnicas de desenho clássicas apresentadas anteriormente estabelecem que podem desenhar-se inversores CMOS com tempos de propagação equilibrados para uma arbitrária capacidade de carga C_L. No entanto, os estudos feitos a este nível [9, 36] mostram que o efeito de C_L, ou se se quiser do *fan-out* da porta, não é negligenciável e contribui para o desequilíbrio entre os tempos de propagação t_{pHL} e t_{pLH}.

As considerações anteriores foram feitas para um sinal de entrada do inversor a comutar instantaneamente entre 0 e V_{DD}, o que implica que apenas um dos transístores esteja activo durante cada uma das transições. Todavia, na realidade o sinal de entrada vai alterando-se gradualmente e, por instantes, o PMOS e o NMOS conduzem em simultâneo, o que acaba por contribuir para que o desequilíbrio entre os tempos de propagação dependa igualmente do tempo de transição da entrada do circuito.

Numa perspectiva de desenho, é complicado descrever a relação entre o tempo de propagação e a transição da entrada. É importante deixar claro que, inserido num circuito digital, um inversor nunca tem uma forma de onda de entrada que seja uma rampa, nem tão pouco um degrau, mas sim a forma de onda de saída da porta que o precede. Significa isto que uma porta nunca deve ser desenhada como se fosse actuar sozinha, pois a sua performance é afectada quer pela capacidade de *drive* da porta que a precede, quer pelo seu

fan-out. Os estudos de Hedenstierna e Jeppson [9] viabilizaram o primeiro modelo que inclui o efeito da transição de entrada do inversor e postulam que um inversor *i*, inserido numa cadeia de inversores, tem o tempo de propagação de

$$t_p^i = t_{step}^i + \eta\, t_{step}^{i-1} \tag{3.12}$$

sendo que o tempo de propagação do inversor, para um degrau de entrada t^i_{step}, vem afectado de uma fracção empírica η do atraso da porta que o precede [9, 29].

3.2.3 – Influência da capacidade de Miller

Outro dos factores que tem impacto ao nível das técnicas de desenho é o efeito de Miller transportado para o inversor. A capacidade de Miller C_M forma-se entre a entrada e a saída, devido às capacidades porta-dreno C_{GD} dos transístores PMOS e NMOS (Figura 3.5). Esta capacidade não deve ser negligenciada pois contribui para o desequilíbrio entre os tempos de propagação do inversor, uma vez que é carregada e descarregada a cada transição do sinal à saída [16, 19].

A capacidade C_{GD} do MOSFET depende, porém, da região de funcionamento em que o transístor se encontra e da tensão aos terminais do dispositivo. Desse modo, a capacidade C_M está igualmente dependente do sinal de entrada do inversor e pode estar sujeita a variações significativas, uma vez que os transístores PMOS e NMOS podem encontrar-se, para cada transição, em diferentes regiões de operação [16]. Na perspectiva dos atrasos de propagação, esta consideração leva a que a capacidade de Miller vá modificar a relação entre os tempos t_{pHL} e t_{pLH} do inversor.

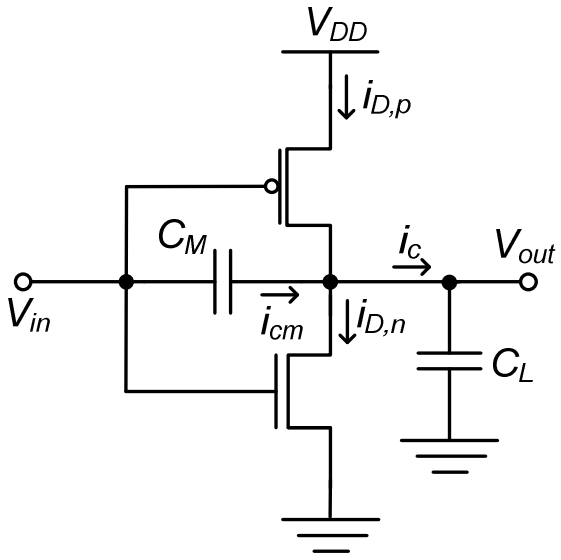

Figura 3.5 – Inversor CMOS com capacidade de Miller entre a entrada e a saída.

Devido ao facto de, na realidade, um circuito integrado real não ter uma entrada do tipo degrau, o PMOS e o NMOS podem encontrar-se no estado activo

simultaneamente. Nesse caso, as correntes no inversor com uma capacidade C_M entre a entrada e a saída e com uma capacidade de carga C_L são as que vêm sinalizadas na Figura 3.5. Os transístores e a capacidade de Miller contribuem para a seguinte equação diferencial do nó de saída

$$i_c(t) = C_L \cdot \frac{dV_{out}}{dt} = i_{cm} + i_{D,p} - i_{D,n} = C_M \left(\frac{dV_{in}}{dt} - \frac{dV_{out}}{dt} \right) + i_{D,p} - i_{D,n} \qquad (3.13)$$

onde i_{Dp} e i_{Dn} são as correntes de dreno do PMOS e do NMOS, respectivamente, e i_{cm} é a corrente que carrega e descarrega a capacidade de Miller entre a entrada e a saída do inversor.

Nas tecnologias actuais, o impacto da capacidade de Miller nas técnicas de desenho de circuitos deve então ser considerado no projecto de circuitos e originou o aparecimento de diversos modelos analíticos. Jeppson foi o primeiro a apresentar um novo modelo [16], sugerindo para o cálculo do t_{pHL} do inversor a seguinte expressão

$$t_{pHL} = \frac{1+2n}{6} t + \frac{\Delta Q}{I_n} + \frac{C_L V_{DD}}{I_n} \left(\frac{C_M}{C_M + C_L} + f(V_{Dsat}) \right) \qquad (3.14)$$

em que t é o tempo de transição da entrada, ΔQ é a carga no nó de saída para a qual o PMOS contribui, I_n a corrente de saturação do NMOS, $f(V_{Dsat})$ uma função da tensão de saturação dreno-fonte e $n = V_{THn} / V_{DD}$ [16]. A equação (3.14) faculta um modelo intuitivo da influência que a capacidade de Miller tem ao nível do desequilíbrio entre os tempos de propagação. A razão C_M/C_L, entre as capacidades do inversor, ganha relevância neste ponto. Para uma razão C_M/C_L pequena, reduz-se o efeito de Miller. Além disso, os estudos feitos por Friedman *et. al* [36] mostram que o efeito da capacidade C_M é menor para tempos de transição de entrada maiores e tem maior impacto nas transições mais rápidas, como Jeppson tinha estudado [16].

3.3 – Circuitos Combinatórios

Adicionando transístores em série e em paralelo a uma topologia do inversor, obtêm-se circuitos de lógica combinatória CMOS. Estes circuitos são a extensão do inversor estático para múltiplas entradas e são compostos por dois circuitos complementares distintos: uma rede só de transístores de tipo P (PUN, *pull-up network*) e por uma rede só de transístores

de tipo N (PDN, *pull-down network*). São estas duas redes que, combinadas, realizam a função lógica que relaciona a saída com as entradas de uma porta CMOS [33].

Para mais facilmente se perceber a natureza particular de um circuito combinatório, atente-se na Figura 3.6. O PUN é responsável pela ligação entre a saída e V_{DD}, para todas as combinações das entradas em que a saída F assume o valor lógico "1". O PDN é activado, providenciando uma ligação entre a saída e a massa, para todas combinações das entradas em que F é "0". Desta relação entre entradas e saída, conclui-se que o PUN e o PDN são mutuamente exclusivos, pois apenas uma das redes é activada de cada vez. Com base nos teoremas de De Morgan, pode demonstrar-se que o PDN e o PUN são circuitos eléctricos duais entre si. A dualidade das portas CMOS diz-nos que a série dos transístores NMOS no PDN corresponde ao paralelo dos PMOS do PUN, e vice-versa [29], sendo esta uma das regras básicas para a construção de circuitos combinatórios. A ligação de transístores NMOS em série realiza a função NAND, pois só se forma um caminho efectivo caso ambas as entradas estiverem a um nível lógico alto, enquanto o paralelo de NMOS corresponde à função NOR, uma vez que existe caminho se pelo menos uma das entradas estiver a "1". Pensando da forma análoga, para o PMOS, a série dos transístores faculta a função NOR, ao passo que o paralelo de PMOS implementa a operação NAND.

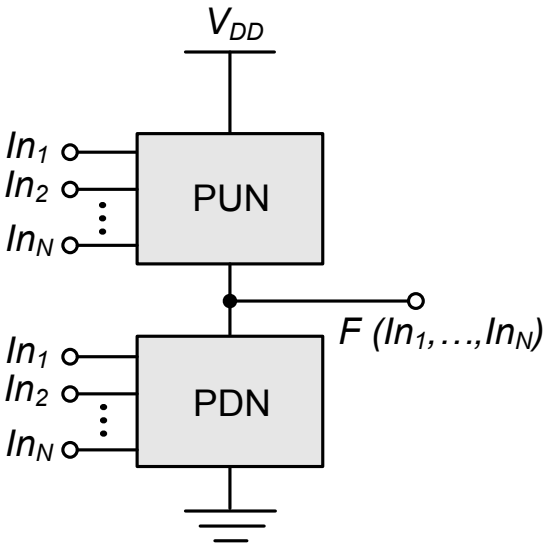

Figura 3.6 – Constituição de um circuito combinatório.

Pode-se, então, construir um circuito combinatório arbitrário começando ora pelo PUN ora pelo PDN, combinando transístores em série ou em paralelo. A outra rede de transístores será sempre obtida aplicando os princípios de dualidade e a junção das duas redes resulta na porta completa. Uma porta CMOS de carácter combinatório é naturalmente inversora, pois precisaria de um segundo andar, com um inversor na saída, para ser não-inversora. Por último, convém referir que as portas CMOS herdam as propriedades do inversor estático: potência estática é praticamente nula (porque não existe,

para DC, um caminho entre a massa e V_{DD}) e os níveis alto e baixo na saída correspondem ao V_{DD} e ao *GND*, respectivamente [29].

3.3.1 – Construção do modelo baseado no inversor equivalente

Por extensão da análise feita para o inversor, pode pensar-se da mesma forma para o desenho de portas lógicas com múltiplas entradas, que implementem outras funções mais complexas. A performance das portas mais simples pode ser estimada através da construção do chamado "inversor equivalente" [15, 19]. Por definição, o inversor equivalente é um circuito que reflecte o comportamento da porta completa e torna mais simples o seu estudo. O dimensionamento das portas lógicas com múltiplas entradas tem, portanto, sempre o inversor equivalente como referência.

Para explicar em que consiste o modelo do inversor equivalente deve ter-se como base as regras para o dimensionamento, por exemplo, das portas lógicas NAND e NOR de duas entradas. Estas regras tradicionais dizem que devem equilibrar-se as resistências equivalentes dos piores percursos do PUN e do PDN, com o compromisso da característica de transferência (VTC) da porta esteja o mais centrada possível e, principalmente, com o compromisso de se obterem tempos de propagação, no pior caso, aproximadamente iguais aos do inversor tido como referência para o dimensionamento. Para que isso se verifique, a resistência associada ao PUN, R_{eqPUN}, deve ser sempre menor ou igual que a resistência do PMOS do inversor de referência e deve respeitar-se a mesma condição para o PDN relativamente ao NMOS do inversor.

Porta NOR

No caso da porta NOR (Figura 3.7), são necessários quatro transístores: dois em paralelo, no PDN, e dois associados em série, no PUN. Se as entradas A e B da porta tiverem valores de tensão inferiores a V_{THn}, os transístores do PDN

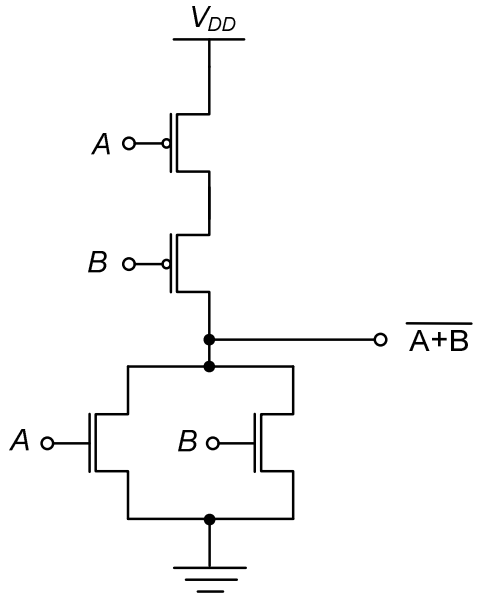

Figura 3.7 – Porta NOR de duas entradas.

estão ao corte (não entram na análise) e os transístores do PUN conduzem. Esta condição leva a que a associação em série das resistências equivalentes dos transístores do PUN aumente a resistência total desse percurso. Dessa forma, conclui-se que esses transístores devam ter o dobro da largura do PMOS do inversor, para salvaguardar o pior caso para o PUN.

$$R_{eqPUN} \propto \frac{1}{k'_p \left(\frac{W}{L}\right)_P} + \frac{1}{k'_p \left(\frac{W}{L}\right)_P} \tag{3.15}$$

Portanto, para transístores com as mesmas dimensões, a resistência equivalente do PUN de uma porta NOR de duas entradas é o dobro da resistência equivalente de um só transístor PMOS do inversor de referência, como vem na expressão (3.15). Ou seja, como para os transístores do PUN o k'_p é constante, a técnica tradicional de desenho diz-nos que o dimensionamento do circuito deve ser respeitar a condição (3.16), com os cálculos a serem feitos interpretando a razão *W/L* como a condutância do transístor. Caso se pretenda, por exemplo, dimensionar uma porta NOR com base num inversor de referência com um transístor PMOS de dimensões (*W/L*) = 0.560 μm/0.13 μm, tem que se construir um inversor equivalente que cumpra esses requisitos, isto é, pela equação (3.16), com os dois PMOS de dimensões (*W/L*) = 1.12 μm/0.13 μm.

$$\left(\frac{W}{L}\right)_{eqPUN} = \frac{1}{\frac{1}{\left(\frac{W}{L}\right)_P} + \frac{1}{\left(\frac{W}{L}\right)_P}} \tag{3.16}$$

Já para a construção do PDN da porta NOR, tem-se que, no pior caso, só um dos transístores NMOS está no estado activo. Na pior das hipóteses, o percurso do PDN é composto por um só transístor. Esta circunstância leva a que a resistência equivalente do PDN seja igual à resistência de um NMOS individual. Logo, tem-se que os transístores do PDN devem ser dimensionados de acordo com a condição (3.17).

$$\frac{1}{\left(\frac{W}{L}\right)_{eqPDN}} = \frac{1}{\left(\frac{W}{L}\right)_N} \quad \Leftrightarrow \quad \left(\frac{W}{L}\right)_{eqPDN} = \left(\frac{W}{L}\right)_N \tag{3.17}$$

Porta NAND

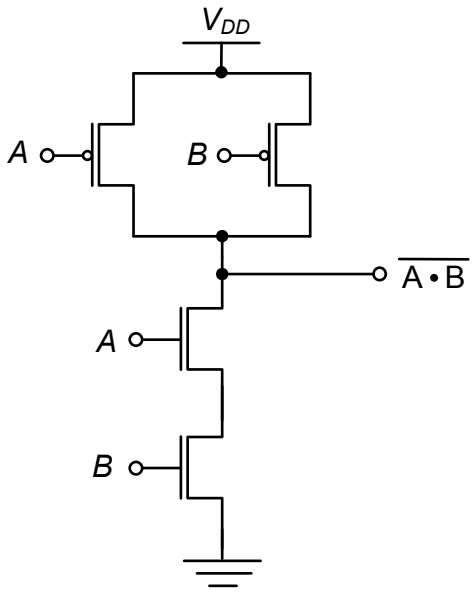

Figura 3.8 – Porta NAND de duas entradas.

Para a porta NAND de duas entradas (Figura 3.8), o PDN é formado pela associação em série de dois transístores NMOS, enquanto o paralelo de dois PMOS constitui o PUN. De forma análoga à análise feita para a NOR, se ambas as entradas da porta activarem a série de transístores do PDN, a resistência do PDN deve ser o dobro da apresentada por um só NMOS. Assim, o dimensionamento do PDN da porta lógica deve ser feito de forma a verificar a equação (3.18). Pretendendo-se uma NAND que tenha por referência um inversor com um NMOS de (W/L) = 0.16 μm/0.13 μm, os transístores que do seu PDN devem ter as dimensões (W/L) = 0.32 μm/0.13 μm.

$$\left(\frac{W}{L}\right)_{eqPDN} = \frac{1}{\frac{1}{\left(\frac{W}{L}\right)_N} + \frac{1}{\left(\frac{W}{L}\right)_N}} \tag{3.18}$$

Quanto ao PUN, que é formado por dois PMOS em paralelo, sabe-se que, no pior caso, só um deles se encontra no estado activo. Esta condição implica que a resistência equivalente do PUN tenha que ser igual à resistência equivalente de um só transístor, pelo que o PUN deve ser construído de modo a que o seu W/L efectivo seja igual do ao PMOS do inversor de referência – equação (3.19).

$$\frac{1}{\left(\frac{W}{L}\right)_{eqPUN}} = \frac{1}{\left(\frac{W}{L}\right)_P} \quad \Leftrightarrow \quad \left(\frac{W}{L}\right)_{eqPUN} = \left(\frac{W}{L}\right)_P \tag{3.19}$$

Comparando as portas NAND e NOR, a NAND é a topologia preferida, apesar de ambas serem largamente utilizadas no projecto de circuitos digitais [19]. Para um igual número de entradas e dimensão dos transístores, a porta NAND apresenta uma melhor resposta transitória (menores atrasos de propagação), o que a torna mais popular no desenho de circuitos. Além disso, ocupa menor área de silício e a associação em série de PMOS deve ser evitada tanto quando possível, por causa das diferenças na mobilidade

efectiva dos portadores de um e de outro dispositivo [23]. O estudo realizado por Kung e Puri [22] faculta um modelo capaz de compensar estas divergências, propondo aproximações diferentes para as estruturas NAND e NOR.

3.4 – Sumário

Neste capítulo apresentaram-se, numa primeira parte, as características mais importantes do inversor CMOS, orientadas para o estudo das principais técnicas de desenho de circuitos digitais. Posteriormente, foi dada maior atenção à discussão das técnicas convencionais de desenho. Explorou-se de forma conveniente o modelo baseado no equilíbrio entre as resistências equivalentes e apresentaram-se ainda outras derivações a esse modelo, estudando igualmente a importância e a influência do *fan-out*, da transição na entrada e da capacidade de Miller entre a entrada e a saída do inversor.

Posteriormente, foi feita uma extensão da análise do inversor para as portas lógicas de múltiplas entradas. Foram discutidas as tradicionais técnicas de desenho, segundo as quais o dimensionamento do PUN e do PDN é feito com base no equilíbrio das resistências dos piores percursos. No que se refere às portas de múltiplas entradas, foram comparadas duas das principais estruturas CMOS: as portas NAND e NOR.

Capítulo 4

Caracterização das portas lógicas

O rápido desenvolvimento das tecnologias CMOS tem vindo a causar inúmeros problemas aos projectistas de circuitos digitais. Ao mesmo tempo, tem criado a necessidade de uma melhor avaliação das tendências tecnológicas, proporcionando desafios ao nível do desenho de circuitos que apenas podem ser estudados mediante a utilização de avançadas ferramentas de CAD (*Computer-Aided Design*). Neste capítulo, numa primeira fase, é descrito todo o trabalho de desenvolvimento das portas lógicas, feito utilizando o ambiente integrado da ferramenta *Cadence DFII*. São ainda apresentadas as opções tomadas ao longo do trabalho, sendo depois feita uma caracterização dos circuitos desenhados e uma consequente apreciação dos resultados extraídos das diversas simulações esquemáticas realizadas.

4.1 – Tecnologias CMOS utilizadas

Ao longo deste trabalho, foram utilizadas cinco tecnologias CMOS diferentes no desenvolvimento das portas lógicas. O estudo inclui duas tecnologias da AMS (*Austria Microsystems*), de 800 nm e 350 nm, porém foi dada maior importância às tecnologias mais recentes, disponibilizadas pela UMC (*United Microelectronics Corporation*), as de 180 nm, 130 nm e 90 nm.

A tecnologia da UMC de 180 nm caracteriza-se por permitir a criação de células com um nível de polisilício e até oito níveis de camadas de metal. As dimensões mínimas de cada MOSFET são W_{min} = 240 nm e L_{min} = 180 nm, sendo que, neste trabalho, optou-se por manter o comprimento mínimo do canal para todos os transístores das portas lógicas desenvolvidas dentro desta tecnologia. Os transístores suportam uma tensão de alimentação de 1.8 V.

O *design-kit* da UMC de 130 nm já possibilita a utilização de transístores de dimensões mínimas de W_{min} = 160 nm e L_{min} = 120 nm, tendo-se utilizado o comprimento mínimo L_{min} de 130 nm. Os transístores utilizados suportam uma tensão de alimentação de 1.2 V. Tal como a tecnologia de 180 nm, a tecnologia de 130 nm também admite o desenvolvimento de células com uma camada de polisilício e com até oito camadas de metal, tendo ainda a opção de metal com duas camadas de espessura diferentes do metal.

A tecnologia da UMC de 90 nm disponibiliza com o seu *design-kit* uma opção que suporta a criação de células que tenham até nove camadas de metal. Os transístores utilizados no trabalho têm dimensões mínimas de W_{min} = 120 nm e L_{min} = 80 nm, mas estabeleceu-se para todas as portas desenvolvidas dentro desta tecnologia um L_{min} de 90 nm. A tensão de alimentação para cada MOSFET é de 1.0 V.

Quanto às duas tecnologias da AMS, a de 800 nm permite que se utilizem transístores de W_{min} = 0.8 μm e L_{min} = 0.8 μm, que suportam uma tensão de alimentação máxima de 5 V. A AMS de 350 nm, por sua vez, possibilita a utilização de transístores com as dimensões mínimas de W_{min} = 0.4 μm e L_{min} = 0.35 μm e os seus modelos típicos suportam uma tensão de 3.3 V.

4.2 – Método de desenvolvimento

Sucintamente, a metodologia de desenvolvimento das portas lógicas segue o diagrama da Figura 4.1 e divide-se em duas fases: desenho esquemático, numa primeira, e posterior implementação em *layout*. A criação do esquemático, utilizando o *Composer Schematic Editor*, constitui o primeiro passo para o desenho de qualquer porta lógica, ao qual se segue a simulação esquemática e extracção da *netlist* do circuito. Se os resultados obtidos nessa primeira simulação estiverem de acordo com as especificações iniciais, inicia-se, através da ferramenta *Virtuoso Layout Editor*, o desenho do *layout*, que assenta num

projecto com base em *standard cells* – isto é, todas as células criadas têm a mesma altura e a sua largura varia consoante a complexidade da porta.

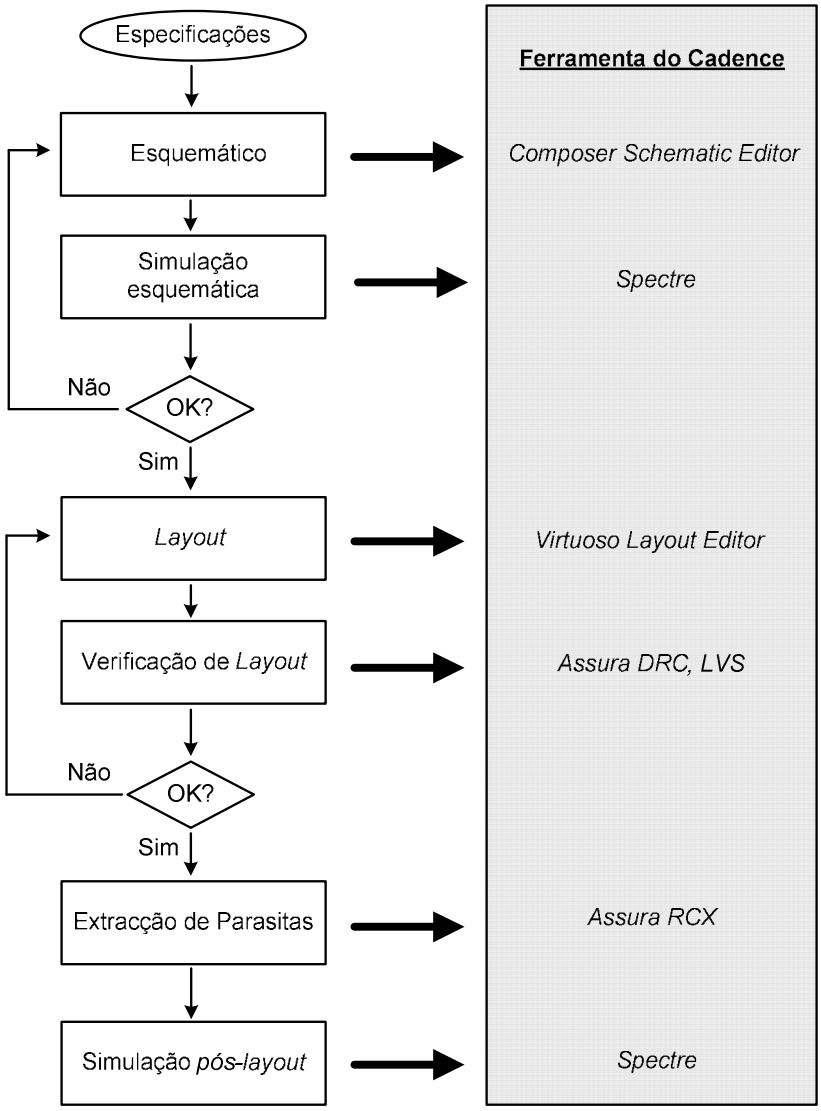

Figura 4.1 – Utilização do Cadence como ferramenta desenvolvimento.

A verificação do desenho é realizada pelo *Assura*, através do DRC (*Design Rule Check*), que certifica se as regras de desenho impostas pela tecnologia são cumpridas, e do LVS (*Layout Vs. Schematic*), para comparar a implementação em *layout* com o esquemático e assegurar que essas duas vistas têm correspondência entre si. Posteriormente, depois de terem passado as ferramentas de verificação, é feita a extracção de capacidades e/ou resistências parasitas pelo módulo RCX (*Resistance/Capacitance and Inductance Extraction*) do *Assura*. A vista de *layout* extraída deste último passo do

desenvolvimento é então utilizada nos trabalhos de simulação *pós-layout*, pelo que a caracterização das portas desenhadas já inclui os elementos parasitas do desenho. No caso deste trabalho, inclui-se apenas a extracção de capacidades parasitas.

4.3 – Estratégias utilizadas no desenho das portas lógicas

O trabalho ao nível do desenho de circuitos de electrónica digital consistiu na criação de um vasto conjunto de portas lógicas CMOS estáticas. O principal objectivo desta dissertação visa a caracterização de uma série de portas lógicas quanto ao impacto que o *scaling* tem nas tradicionais técnicas de desenho de circuitos, pelo que foram alvo de estudo diversas tecnologias CMOS de comprimento de canal diferente.

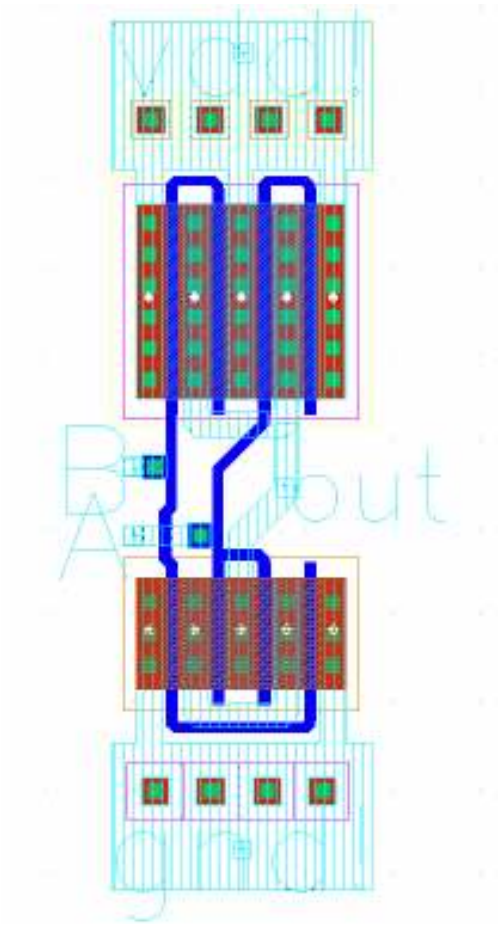

Figura 4.2 – Vista de desenho de *layout*.

Para se ter uma base de comparação apropriada entre as tecnologias, estabeleceu-se que, para cada uma, seriam desenvolvidas uma série de portas lógicas com diferentes composições do PUN e do PDN, por forma a indexar as estruturas CMOS mais relevantes e, ao mesmo tempo, explorar convenientemente o efeito que as dimensões dos transístores têm na aplicação das técnicas de desenho. De salientar que a principal premissa para a criação das células para cada tecnologia passa por respeitar o factor de *scaling* entre cada uma delas. Por exemplo, da tecnologia de 180 nm para a tecnologia de 130 nm, as células criadas deverão respeitar o factor de *scaling* entre tecnologias ao nível de todas as dimensões físicas da célula, desde as dimensões dos transístores, até à largura das ligações metálicas ou altura das células. Na Figura 4.2 vem o exemplo de um desenho de *layout*, no caso uma porta NAND de duas entradas.

Neste trabalho, optou-se pelo desenho de um conjunto predefinido de portas lógicas estáticas. Estabeleceu-se que, para cada uma das tecnologias, inicialmente seria feito o desenho e dimensionamento de quatro inversores estáticos com vários tamanhos (com diferentes larguras do transístor PMOS e do NMOS), mas mantendo constante a razão entre as dimensões dos transístores. Posteriormente, com base na construção do modelo baseado no inversor equivalente e tendo por base os quatro inversores de referência dimensionados, estas considerações foram estendidas para o desenho de portas com múltiplas entradas, entre portas NAND e NOR de duas e três entradas, também elas com quatro versões distintas de *fan-out*[2].

Para cada uma das tecnologias em estudo resulta que foram, portanto, desenvolvidas 20 portas lógicas estáticas, que passaram a fazer parte de uma biblioteca de componentes para simulação. Advém daí a necessidade de definir um conjunto de testes que facilitem o aturado trabalho de simulação e de caracterização das portas. Dentro do *Cadence*, utilizou-se o simulador *Spectre* e definiram-se diversos ambientes de teste e várias configurações no *Analog Design Environment*, de modo a tornar mais rápido o processo de obtenção de resultados.

A Figura 4.3 mostra o ambiente de teste de um inversor: um esquemático de *fan-out* unitário, com um inversor a fazer o *drive* de outro e uma capacidade mínima na saída do segundo inversor. O esquema foi replicado para todas as portas e em todas as tecnologias foi utilizado, numa primeira caracterização, estímulo do circuito do tipo degrau, de amplitude entre 0 e V_{DD}, com um tempo subida/descida de transição da entrada de 1 ps.

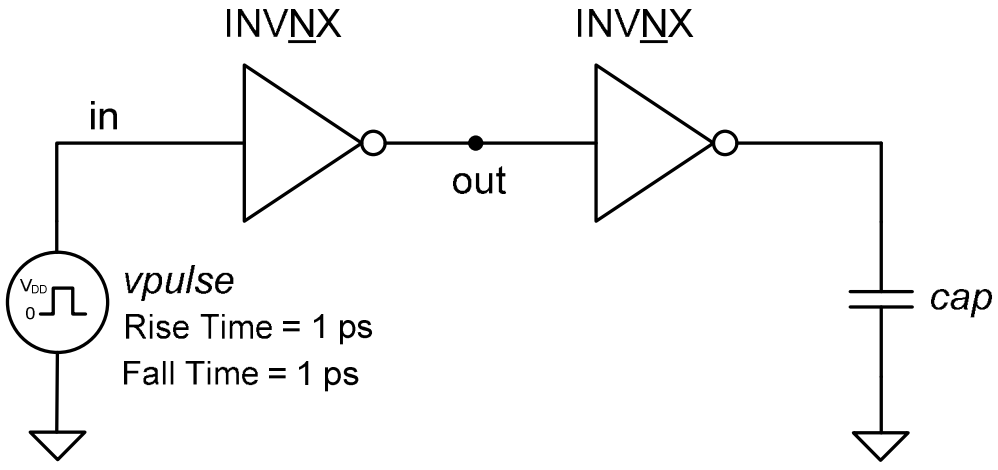

Figura 4.3 – Esquemático do ambiente de teste de um inversor estático.

[2] ***fan-out*** – define-se como o número de portas ligadas na saída de uma porta *driver*.

4.4 – Resultados

O trabalho de simulação das portas lógicas desenvolvidas incidiu em dois pontos: na caracterização das técnicas tradicionais de desenho para o dimensionamento de um conjunto de inversores de vários tamanhos; e no estudo dessas técnicas para o caso da construção de portas com mais entradas. Neste ponto introdutório, convém deixar bem claro que optou-se por fazer a separação entre os dois tipos de abordagem: os estudos feitos ao inversor de um lado, os estudos realizados às portas NAND e NOR de outro.

4.4.1 – Caracterização do inversor estático

Como se pretende avaliar as regras tradicionais de desenho utilizadas no dimensionamento do inversor, estabeleceu-se que essa avaliação seria quantificada por um parâmetro que se designou por δ. O parâmetro δ, expresso por (4.1), é tido como o desequilíbrio percentual entre os tempos de propagação do circuito e constituiu um elemento de comparação entre as diversas tecnologias que o trabalho abrange.

$$\delta = \frac{t_{pHL} - t_{pLH}}{t_p} \times 100\,(\%) \tag{4.1}$$

4.4.1.1 – Estudo isolado de uma tecnologia

Destacada esta premissa, no trabalho de desenvolvimento e caracterização das portas lógicas começou por centrar-se atenções apenas numa das tecnologias alvo de estudo: a UMC de 130 nm. Desenharam-se uma série de inversores utilizando transístores desta tecnologia, sendo que o dimensionamento foi feito através do ajuste dos parâmetros de desenho W_p e W_n de forma a, numa primeira fase, compensar apenas as diferenças entre a mobilidade dos portadores do NMOS e a do PMOS, ou seja:

$$\beta = \frac{\mu_n}{\mu_p} = \frac{W_p}{W_n} \tag{4.2}$$

Para a UMC130 foram então desenhados quatro inversores com diversos tamanhos. O inversor 1X foi construído utilizando um transístor NMOS com dimensões mínimas (W_{min} = 160 nm), enquanto ao PMOS foi aplicada a razão de desenho. A nomenclatura 1X

corresponde, portanto, ao inversor de tamanho menor. Assentando na suposição que um inversor com o dobro da largura dos seus transístores apresenta uma condutância eléctrica duas vezes superior, foram desenhados outros três inversores com base no inversor de dimensões mínimas: 2X, 4X e 8X. O inversor 8X corresponde, claro está, a um inversor composto por transístores com larguras oito vezes superiores às do inversor 1X.

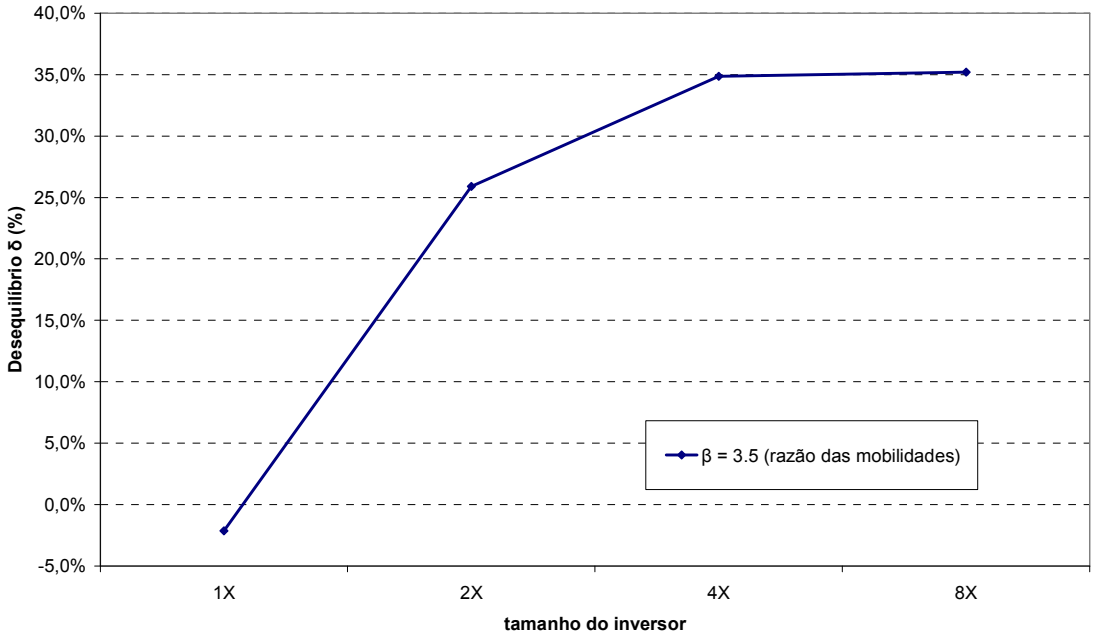

Figura 4.4 – UMC130: desequilíbrio δ em função do tamanho do inversor para a técnica da razão das mobilidades.

Dimensionados os inversores dessa tecnologia, o trabalho de simulação conduz ao gráfico de resultados da Figura 4.4, que relaciona o desequilíbrio entre os tempos de propagação δ com o tamanho dos transístores do inversor. Daqui, verifica-se que o desequilíbrio δ aumenta para inversores desenhados com inversores de maiores dimensões. O que contraria o que se esperava para estes resultados, pois era de prever que, para inversores desenhados com transístores de dimensões mais reduzidas, o modelo convencional de desenho fosse originando divergências cada vez mais notórias entre os atrasos de propagação. Os resultados obtidos, pelo contrário, mostram que o inversor construído com transístores de dimensões mínimas apresenta uma característica de transferência de elevada simetria, com tempos de propagação equilibrados. O aumento do tamanho do inversor corresponde a um aumento do desequilíbrio δ entre os tempos de propagação. Nota-se que, à medida que o tamanho do inversor aumenta, os tempos de propagação diminuem, mas de maneira diferente: o t_{pHL} diminui de forma moderada,

enquanto o t_{pLH} diminui de uma forma mais acentuada, o que dá origem a desequilíbrios mais significativos.

Há, portanto, uma série de factores com influência na técnica convencional de desenho. Relembre-se que, por exemplo, a técnica tradicional de desenho assenta na compensação das mobilidades dos transístores se se considerar que as tensões de limiar dos mesmos são iguais. Todavia, é sabido que na prática o PMOS e o NMOS não são criados com tensões V_{TH} iguais. Deste modo, no sentido de começar a deslindar alguns desses factores, realizaram-se novos testes de simulação utilizando diferentes razões entre a largura dos transístores W_p/W_n. Redesenharam-se então os inversores da UMC130 utilizando outro modelo analítico, o da equação (4.3), que entra também com as desigualdades entre as tensões de limiar V_{THp} e V_{THn}.

$$\beta = \frac{W_p}{W_n} = \frac{(V_{DD} - V_{THn})}{(V_{DD} - |V_{THp}|)} \cdot \frac{\left[\frac{2\,|V_{THp}|}{V_{DD} - |V_{THp}|} + \ln\left(\frac{4\,(V_{DD} - |V_{THp}|)}{V_{DD}} - 1\right)\right]}{\left[\frac{2\,V_{THn}}{V_{DD} - V_{THn}} + \ln\left(\frac{4\,(V_{DD} - V_{THn})}{V_{DD}} - 1\right)\right]} \cdot \frac{\mu_n}{\mu_p} \qquad (4.3)$$

Adicionando o novo β de desenho à análise obtiveram-se resultados em tudo semelhantes aos anteriores. Dessa forma, optou-se por adicionar outras duas razões de desenho: uma obtida empiricamente e outra para um β de desenho intermédio. Por simulação das vistas esquemáticas, experimentaram-se vários valores para a razão W_p/W_n que, para os inversores de tamanhos maiores, implicava um desequilíbrio pouco significativo. Optou-se por aplicar o ajuste empírico aos inversores de tamanhos maiores porque notou-se, no estudo anterior, que para estes inversores o desequilíbrio δ se apresentava praticamente independente das dimensões dos transístores. Da Figura 4.5, conclui-se que o modelo da equação (4.3) conduz ainda a desequilíbrios maiores entre os tempos de propagação, o que indica que as diferenças entre as tensões de limiar V_{TH} não têm assim tanto impacto a este nível. A razão de desenho intermédia, com um valor de β entre o ajuste empírico e o método da compensação das mobilidades, apresenta-se como a mais sensata das quatro e estabelece uma aproximação a um β óptimo, obtido no entanto por simulação esquemática (com todos os erros inerentes a essa estratégia).

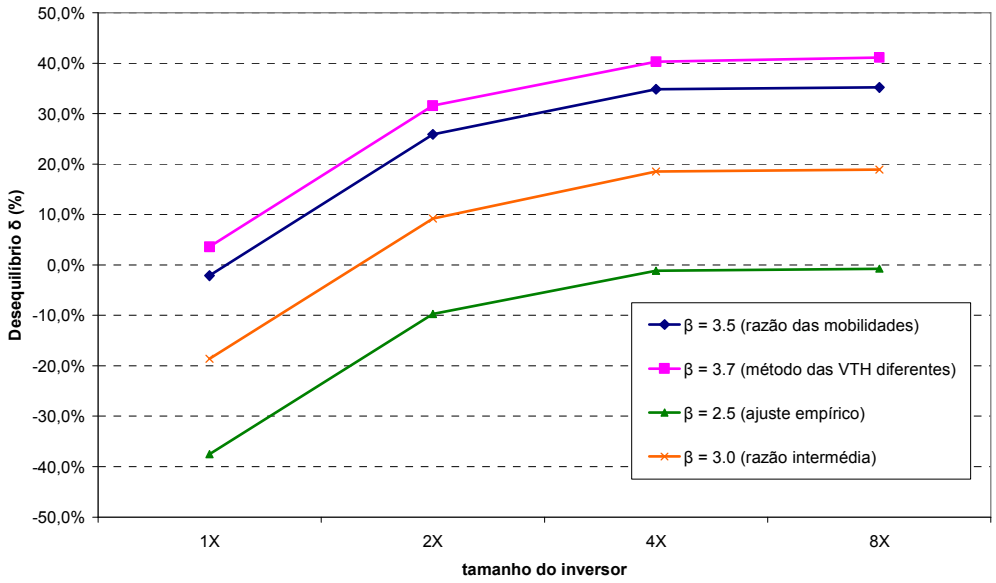

Figura 4.5 – UMC130: influência da razão de desenho.

Recorde-se que estes resultados foram obtidos utilizando o ambiente de teste da Figura 4.3, com *fan-out* unitário e com um estímulo na entrada do tipo degrau – com um tempo de transição de 1 ps. No entanto, esta configuração não é muito realista. Inserido num circuito digital, o inversor tem mais que uma porta ligada à sua saída (o *fan-out* não é unitário) e o seu sinal de entrada não é um sinal do tipo degrau, mas sim uma outra forma de onda com um determinado tempo de transição.

Assim, para o ajuste de desenho da equação (4.3), realizaram-se mais uma série de ensaios para diversos tipos de *fan-out* e para uma gama de valores de transição da entrada. Começou-se então pelo estudo da influência do *fan-out* (N) no desequilíbrio entre os tempos de propagação, mantendo a transição do sinal de entrada do tipo degrau. Realizaram-se testes aos inversores desenhados para as tecnologias UMC130 e obtiveram-se resultados que exprimem o desequilíbrio δ para dois tipos de *fan-out*: $N = 4$ e $N = 8$.

Na Figura 4.6 verifica-se que, para valores de *fan-out* superiores, reduz-se o desequilíbrio δ. A carga que o inversor tem à sua saída para carregar é maior, para $N = 4$ e $N = 8$, o que leva a que o tempo que demora a carregar essa carga seja maior. Desta forma, há um desequilíbrio menor entre o t_{pHL} e o t_{pLH}. Além disso, para um *fan-out* superior, o impacto da capacidade de Miller C_M no desenho é menor. Conclui-se que o facto da razão de capacidades C_M/C_L ser reduzida (C_M muito pequena comparativamente à capacidade de carga C_L) leva a que a influência da C_M seja menor e conduz a um menor desequilíbrio entre os tempos de propagação.

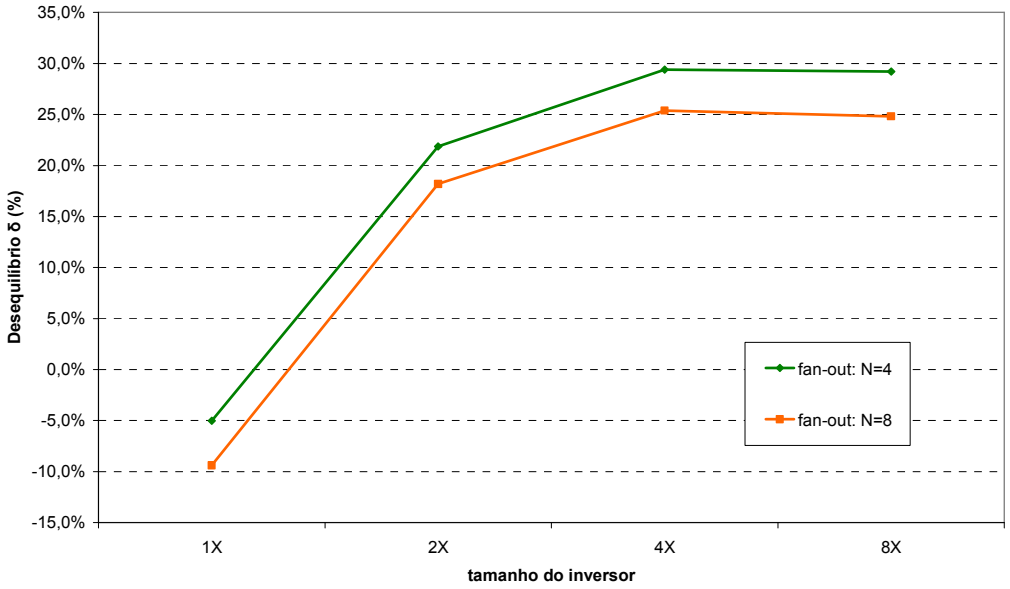

Figura 4.6 – UMC130: desequilíbrio δ em função do *fan-out*.

Noutro âmbito, para avaliar a influência do tempo de transição do estímulo de entrada nos resultados anteriores, realizaram-se igualmente testes para quantificar o desequilíbrio δ em função de dois parâmetros: o *fan-out* e o tempo de transição da entrada, que se designou por *trin*.

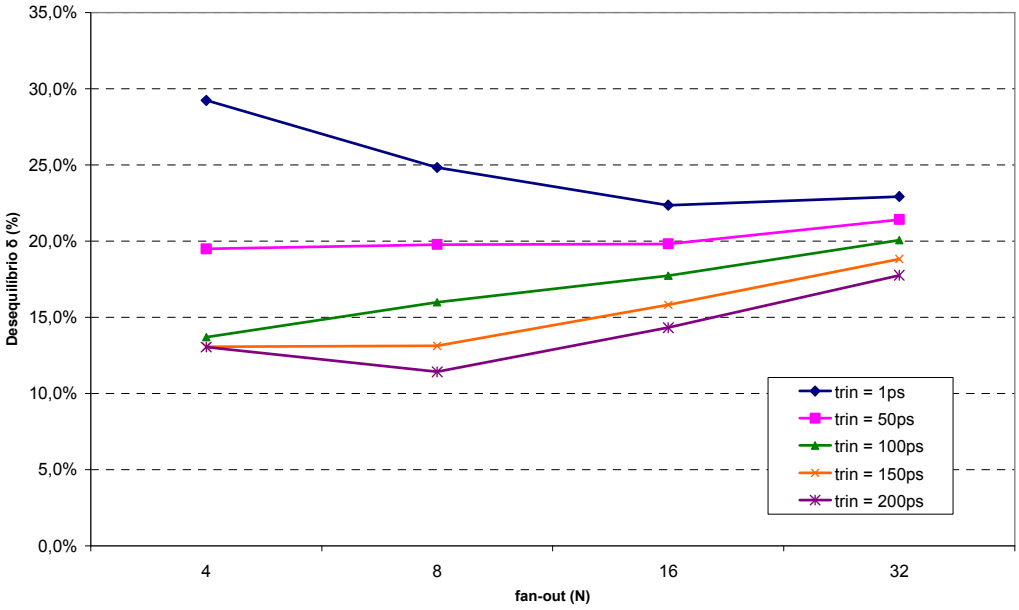

Figura 4.7 – UMC130: influência do *fan-out* e do *trin* no desenho do INV8X.

A Figura 4.7 mostra os resultados do desequilíbrio δ para quatro valores diferentes de *fan-out* e para cinco tipos de transição da entrada, desde o estímulo do tipo degrau (1 ps) até ao do tipo rampa com declive bem menos acentuado (200 ps). Quanto à influência da

transição da entrada nota-se que, à medida que o *fan-out* e o *trin* aumentam, os cinco traçados interpolados vão convergindo, originando diferenças menos significativas entre os desequilíbrios δ. Segundo o modelo apresentado na Secção 3.2.3, estes resultados indicam que o efeito da capacidade de Miller C_M deixa de ter tanto impacto a este nível.

Adicionalmente, comprova-se também que o efeito da capacidade C_M reduz-se linearmente com o aumento do parâmetro *trin*. Para uma transição mais rápida, a derivada do sinal de entrada é maior e a influência da capacidade de Miller torna-se mais significativa, conduzindo a um desequilíbrio δ maior do que para o caso da transição da entrada se fizer de uma forma mais gradual. Está à vista que as diferenças no estímulo da entrada motivam alterações no ritmo com que a capacidade C_M é carregada ou descarregada, o que no final acaba por provocar algumas discrepâncias no que respeita aos tempos de propagação de um circuito.

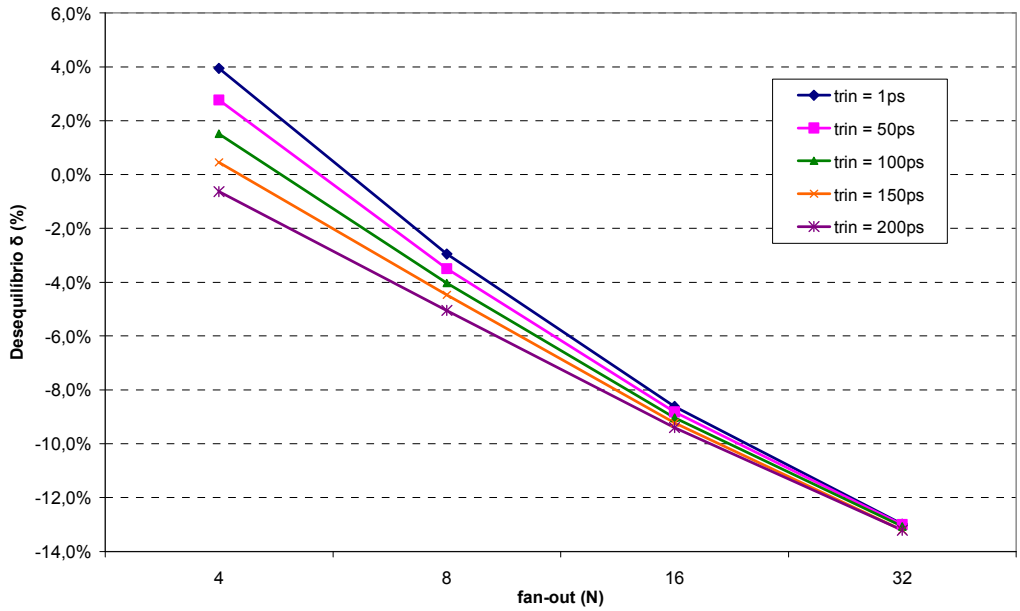

Figura 4.8 – AMS800: influência do *fan-out* e do *trin* no desenho do INV8X.

Para uma tecnologia de dimensões maiores, como a AMS800, a interpretação é ainda mais clara (Figura 4.8). Mantém-se a tendência: para valores de *fan-out* superiores, o desequilíbrio δ do inversor tende a convergir para um valor comum. Para um valor razoável de *fan-out*, como $N = 4$, verifica-se que o *trin* da entrada ainda origina uma maior discrepância entre os desequilíbrios δ, mas essa diferença diminui à medida que o *fan-out* aumenta. Ao nível dos tempos de propagação, conclui-se que o facto do desequilíbrio δ assumir valores cada vez mais negativos deve-se à perda de rendimento do transístor

PMOS. A carga do circuito aumenta, o PMOS demora mais tempo a carregar essa mesma carga, motivando o aumento mais significativo do t_{pLH}, que levará a equação (4.1) para valores negativos. Daqui pode concluir-se que a razão de desenho utilizada não é mais adequada. Ou seja, a técnica (4.3), que inclui no modelo as diferenças entre as tensões de limiar V_{TH}, não é também apropriada, pois provoca uma considerável assimetria no circuito do ponto de vista dos tempos de propagação.

4.4.1.2 – Comparação entre as cinco tecnologias utilizadas

Feito este estudo, evoluiu-se para um estudo mais abrangente, que incluísse resultados de todas as tecnologias do plano de trabalhos. Para cada tecnologia, utilizando a regra de desenho da equação (4.2), foram igualmente dimensionados quatro inversores com diversos tamanhos. O aturado trabalho de simulação conduz ao conjunto de curvas da Figura 4.9, que estabelece um termo de comparação entre as cinco tecnologias utilizadas, exprimindo o desequilíbrio δ em função do tamanho do inversor.

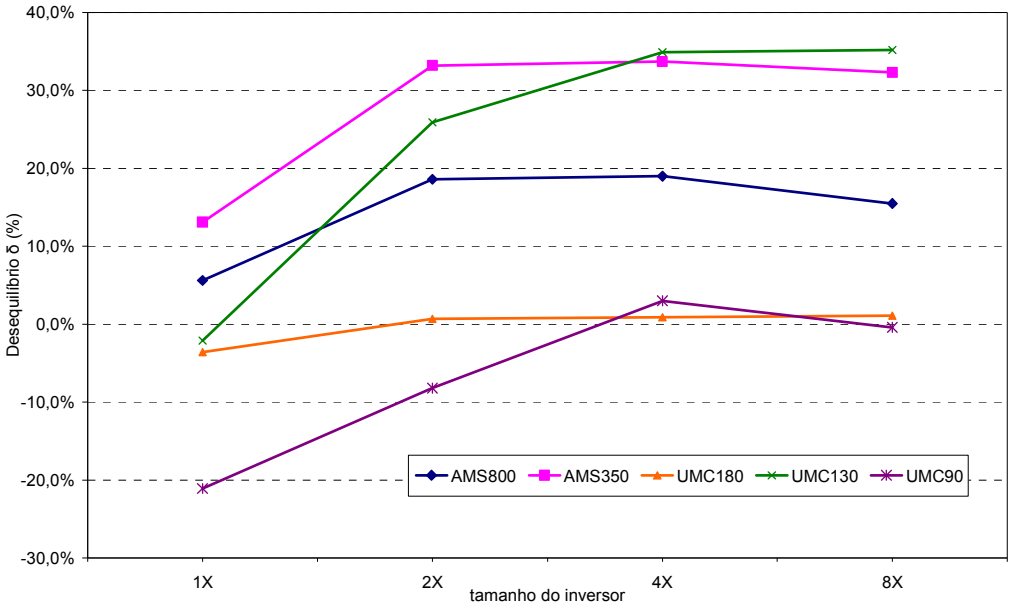

Figura 4.9 – Caracterização das diferentes tecnologias em termos de desequilíbrio δ.

Esperava-se que este estudo mostrasse que, para portas desenhadas com transístores de dimensões cada vez mais reduzidas, o modelo tradicional de desenho fosse implicando divergências cada vez maiores quanto ao desequilíbrio δ. Dos resultados da Figura 4.9, verifica-se que, dentro de cada fabricante, o desequilíbrio característico de cada tecnologia

acentua-se para as tecnologias de dimensões mais reduzidas. Colocando de parte a tecnologia de 90 nm (que apresenta um modelo de simulação mais inconsistente), verifica-se uma perda de desempenho quando se utilizam tecnologias de dimensões mais reduzidas, dentro do mesmo fabricante. Desta forma, é possível relacionar os resultados obtidos directamente com o *scaling* da tecnologia. Todavia, nota-se que o desequilíbrio δ varia de uma maneira consideravelmente irregular com a tecnologia utilizada, o que leva a crer que existam diferenças substanciais entre os modelos de simulação dos transístores de cada processo.

Dentro destes resultados, nota-se que a evolução da tecnologia AMS800 para a AMS350 motivou um pior desempenho da técnica tradicional de desenho, que se revelou menos eficiente para a tecnologia de 350 nm, pois conduz a desequilíbrios maiores entre os atrasos das portas lógicas. No que respeita às tecnologias da UMC, de comprimento de canal mais reduzido, nota-se também que existe uma perda de performance com o *scaling* da tecnologia de 180 nm para a de 130 nm. Confirma-se que nas tecnologias de maiores dimensões (180, 350 e 800 nm) o desequilíbrio δ é praticamente independente do tamanho do inversor, caso se considerem os inversores de tamanhos maiores. É ainda curioso o facto da tecnologia de 180 nm revelar um comportamento melhor que a de 350 nm, pois é caracterizada não só por ter tempos de propagação próximos do equilíbrio, como também por ser praticamente insensível ao tamanho do inversor.

Em termos quantitativos, constata-se o mesmo que já se tinha concluído no estudo particular da tecnologia de 130 nm. E confirma-se que os tempos de propagação t_{pHL} e t_{pLH} evoluem desse mesmo modo para todas as tecnologias alvo de estudo. À medida que se aumentam as dimensões dos transístores, nota-se que os tempos de propagação diminuem, mas com uma pequena diferença: o t_{pHL} diminui moderadamente, enquanto o t_{pLH} diminui de uma forma muito mais acentuada. Isto é, o t_{pLH} mostra-se mais sensível às alterações nas dimensões do PMOS do que o t_{pHL} a alterações da mesma ordem no NMOS.

De seguida, noutro ponto do trabalho, dimensionou-se o inversor através de um ajuste empírico. Por simulação esquemática, experimentaram-se diversos valores para as larguras dos transístores e chegou-se a uma razão W_p/W_n que, para os inversores de tamanhos maiores, conduzia a um desequilíbrio pouco significativo. Repetiu-se o processo para todas as tecnologias propostas e do trabalho de simulação das portas desenhadas extraíram-se os resultados para o gráfico da Figura 4.10.

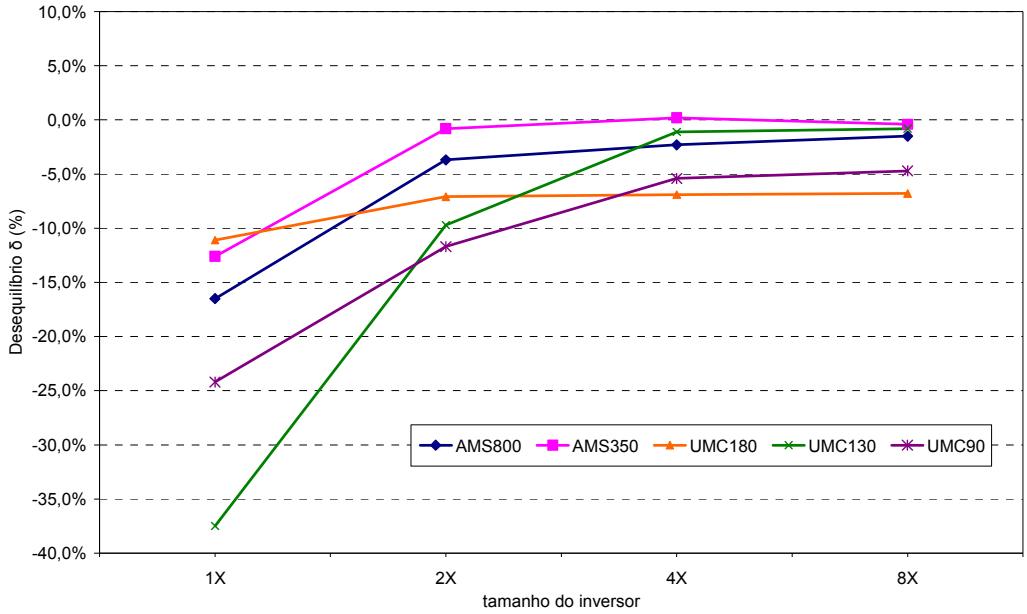

Figura 4.10 – Caracterização das diferentes tecnologias em termos de desequilíbrio δ para o ajuste de desenho empírico.

Os resultados da Figura 4.10 mostram, desde logo, que o ajuste empírico motiva um desequilíbrio entre tempos de propagação reduzido e praticamente constante para os inversores de tamanhos maiores. Não seriam de esperar outros resultados, por causa do ajuste ter sido efectuado para estes casos. Todavia, a redução das dimensões dos transístores que compõem o inversor implica um aumento no desequilíbrio δ. Depois, para os inversores construídos com transístores que tenham as dimensões mínimas, já existe uma discrepância considerável entre os valores de δ. Constata-se também que os tempos de propagação apresentam uma tendência semelhante à do dimensionamento anterior. O t_{pLH} aumenta de maneira bastante pronunciada com a redução do tamanho do inversor e força variações no desequilíbrio δ, que se torna mais negativo. No que respeita ao relacionamento destes resultados com *scaling*, constata-se que os traçados voltam a não seguir um comportamento regular, mas evoluem do mesmo modo para cada fabricante, consoante se utilizem tecnologias da AMS ou da UMC.

As variações verificadas nos resultados anteriores poderão depender de inúmeros factores. Isto porque a razão de desenho pode ser substancialmente afectada por efeitos de canal estreito como a saturação de velocidade de deriva, a degradação da mobilidade ou o DIBL. No entanto, a capacidade de Miller entre a entrada e a saída do inversor é uma das variáveis com maior influência nos resultados e comportamentos obtidos na análise anterior.

4.4.2 – Caracterização das portas NAND e NOR

Noutro ponto do plano de trabalhos, avaliaram-se de igual modo as técnicas de desenho baseadas na construção do modelo do inversor equivalente para o dimensionamento do PUN e do PDN. No caso das portas NAND e NOR, de duas e três entradas, uma vez que estes circuitos são consideravelmente assimétricos, deixa de fazer sentido falar-se no desequilíbrio δ entre os tempos de propagação e centram-se atenções na análise dos piores percursos, comparando os tempos de propagação dos piores casos das várias combinações das entradas da porta lógica com os resultados obtidos para o inversor estático.

UMC130	**Pior Caso**			t_{pinv} **(ps)**	$Erro_{wc}$ **(%)**
	t_{pHL} **(ps)**	t_{pLH} **(ps)**	t_{pwc} **(ps)**		
NAND2_1X	29,58	32,48	**31,03**	**24,70**	*25,7%*
NAND2_2X	24,13	24,46	**24,30**	**19,84**	*22,5%*
NAND2_4X	20,65	20,38	**20,52**	**16,35**	*25,5%*
NAND2_8X	19,82	19,58	**19,70**	**14,32**	*37,6%*
NAND3_1X	35,98	45,05	**40,52**	**24,70**	*48,5%*
NAND3_2X	29,86	36,14	**33,00**	**19,84**	*49,8%*
NAND3_4X	26,39	31,52	**28,96**	**16,35**	*55,7%*
NAND3_8X	24,76	29,77	**27,27**	**14,32**	*62,3%*
NOR2_1X	44,8	35,55	**40,18**	**24,70**	*62,7%*
NOR2_2X	47,71	27,44	**37,58**	**19,84**	*89,4%*
NOR2_4X	43,47	24,29	**33,88**	**16,35**	*107,2%*
NOR2_8X	42,88	23,89	**33,39**	**14,32**	*133,1%*
NOR3_1X	75,6	52,19	**63,90**	**24,70**	*158,7%*
NOR3_2X	86,94	43,24	**65,09**	**19,84**	*228,1%*
NOR3_4X	81,74	39,53	**60,64**	**16,35**	*270,9%*
NOR3_8X	81,11	39,1	**60,11**	**14,32**	*319,7%*

Tabela 4.1 – Análise dos piores casos das portas da tecnologia UMC130.

Para avaliar de forma coerente a técnica de desenho, estabeleceu-se um parâmetro de erro designado por $Erro_{wc}$ para facultar um termo de comparação entre o pior caso das portas NAND e NOR e o tempo de propagação do inversor que essas portas de múltiplas

entradas têm como referência para o seu dimensionamento. A Tabela 4.1, com resultados obtidos para a tecnologia UMC de 130 nm, lista: nas primeiras três colunas de valores, o pior caso da resposta transitória das portas lógicas desenhadas; na quarta coluna, o tempo de propagação do inversor de referência, t_{pinv}; e na coluna mais à direita, o parâmetro de erro entre a análise de piores casos e o tempo de propagação do inversor de referência

Através dos resultados da Tabela 4.1, salta à vista que a técnica tradicional de desenho de portas lógicas conduz a maiores divergências entre os tempos de propagação quando é preciso construir estruturas que recorram a transístores de maiores dimensões. Para a porta NAND de duas entradas, o dimensionamento do PUN e do PDN pelo método convencional, com base nos piores percursos, não conduz a circuitos tão erróneos quanto isso. No entanto, a porta NAND-3, mais complexa, já apresenta um erro maior relativamente ao inversor de referência. Depreende-se então que a performance da porta NAND degrada-se rapidamente com o aumento do *fan-in*, por culpa do aumento das capacidades parasitas associadas aos transístores do PUN e do PDN [29].

Depois, outra das conclusões mais importantes a reter é a da importância de se evitar a associação em série de transístores, como já havia sido referido na Secção 3.3.1. Devido ao facto dos transístores do tipo P terem uma mobilidade dos portadores inferior à dos transístores do tipo N, o *stack* de PMOS motiva piores resultados na resposta transitória da porta. Os dados da Tabela 4.1 comprovam isso mesmo, mostrando que, para as portas NOR, a regra de desenho praticamente deixa de fazer sentido, originando divergências enormes entre o pior caso do PUN e do PDN e o atraso de propagação do inversor de referência.

A associação em série de transístores NMOS já não causa tantos problemas ao nível do tempo de propagação da porta. Constata-se isso mesmo analisando os resultados da porta NAND-2 e comparando os valores obtidos para o t_{pHL}, atraso pelo qual o NMOS é "responsável", com os valores do inversor estático: a diferença entre os tempos é menor ainda e o erro seria bastante diminuto. Também para a NAND-2 é o PMOS que influi negativamente na análise dos piores percursos.

Na Tabela 4.2, que apresenta a análise dos piores casos das portas desenhadas para a tecnologia AMS800, a interpretação dos resultados é semelhante. De forma análoga, conclui-se que a porta NAND é a melhor em termos de performance e a porta para o qual o modelo tradicional de desenho se mantém exequível.

AMS800	**Pior Caso**			t_{pinv} **(ps)**	$Erro_{wc}$ **(%)**
	t_{pHL} **(ps)**	t_{pLH} **(ps)**	t_{pwc} **(ps)**		
NAND2_1X	236,8	303,3	**270,05**	**248,55**	*8,7%*
NAND2_2X	162,2	204,3	**183,25**	**164,15**	*11,6%*
NAND2_4X	126,9	157,8	**142,35**	**121,90**	*16,8%*
NAND2_8X	108,9	135,8	**122,35**	**99,97**	*22,4%*
NAND3_1X	280	435,4	**357,70**	**248,55**	*43,9%*
NAND3_2X	202,3	311,6	**256,95**	**164,15**	*56,5%*
NAND3_4X	166,3	252,9	**209,60**	**121,90**	*71,9%*
NAND3_8X	143,3	222,6	**182,95**	**99,97**	*83,0%*
NOR2_1X	420,7	303	**361,85**	**248,55**	*45,6%*
NOR2_2X	335	214	**274,50**	**164,15**	*67,2%*
NOR2_4X	272,6	176,3	**224,45**	**121,90**	*84,1%*
NOR2_8X	233,1	150,4	**191,75**	**99,97**	*91,8%*
NOR3_1X	690,6	433,1	**561,85**	**248,55**	*126,1%*
NOR3_2X	591,3	329,5	**460,40**	**164,15**	*180,5%*
NOR3_4X	516	280,6	**398,30**	**121,90**	*226,7%*
NOR3_8X	515,2	279,6	**397,40**	**99,97**	*297,5%*

Tabela 4.2 – Análise dos piores casos das portas da tecnologia AMS800.

Para as outras tecnologias verificaram-se resultados em tudo semelhantes, pelo que não se achou relevante para a discussão apresentar aqui as outras tabelas, do trabalho de caracterização feito para as tecnologias UMC180 e AMS350.

4.5 – Sumário

Neste capítulo foi descrito o método de desenvolvimento das portas lógicas desenhadas, que assentou na simulação esquemática e no desenho de *layout*, indicando as estratégias adoptadas ao longo do trabalho. Apresentaram-se os resultados obtidos para a caracterização das portas desenvolvidas e comentaram-se esses mesmos resultados.

Os resultados extraídos do trabalho de simulação mostram que as técnicas de desenho tradicionais já não se adequam ao desenho optimizado de circuitos que utilizem as

tecnologias actuais, uma vez que produzem divergências significativas no que diz respeito aos atrasos de propagação de uma porta lógica. Obtiveram-se, portanto, resultados conclusivos relativamente ao impacto do *scaling* da tecnologia na aplicação dos modelos convencionais de desenho.

Capítulo 5

Conclusões

Neste trabalho de dissertação foram avaliadas as técnicas tradicionais de desenho de circuitos. Sabe-se que a evolução das tecnologias CMOS tem oferecido aos projectistas de circuitos digitais desafios cada vez maiores ao nível do desenho, pelo que têm sido feitos esforços no sentido de encontrar soluções e adoptar novos modelos analíticos. É no sentido de analisar esses modelos que realizou este estudo.

O estudo feito ao longo desta dissertação assemelha-se mais a um trabalho de exploração, sendo que a sua principal contribuição assenta no facto de se terem estudado portas lógicas construídas com transístores de dimensões muito reduzidas. Por isso mesmo, acredita-se que este trabalho pode inferir novas linhas de orientação que facilitem o projecto de circuitos que recorram às tecnologias actuais. Ao comprovar que as técnicas convencionais de desenho se revelam ineficazes no desenho de circuitos das tecnologias correntes, está-se também a dar outras perspectivas do que é trabalhar com transístores de canais curtos.

Analisando o trabalho como um todo, constata-se que este assenta numa boa parte teórica, que foi explorada nos primeiros capítulos desta dissertação. Começou-se por centrar atenções nos fundamentos relativos à tecnologia CMOS e nos conceitos base essenciais à realização de qualquer estudo sobre o comportamento de um dado circuito de electrónica digital. Posteriormente, foram discutidas as principais técnicas tradicionais de desenho de circuitos em CMOS e os modelos analíticos formulados para ultrapassar

algumas das limitações das abordagens convencionais. Foi com ponto de partida nessas considerações que se projectaram e desenvolveram as portas lógicas estáticas.

Como conclusão de todo o trabalho de simulação efectuado e dos resultados obtidos, a comparação entre as tecnologias CMOS utilizadas acabou por conduzir a resultados importantes ao nível do desenho de circuitos das tecnologias de dimensões mais reduzidas. Desde que trabalhando com tecnologias dentro do mesmo fabricante, conseguiram-se relacionar directamente os resultados obtidos com o *scaling* da tecnologia. Apesar do paralelismo entre tecnologias não ter sido exactamente o que se esperava extrair deste estudo, muitos dos resultados obtidos comprovam o que as plataformas teóricas faziam prever, revelando ainda dados que são relevantes para o tema.

Salienta-se ainda a importância de neste trabalho se ter passado por mais fases de desenvolvimento, atingindo a camada física do circuito integrado através do desenho de *layout*. Verifica-se que o projecto de circuitos digitais ganha novos contornos quando não se faz apenas simulação esquemática. Trabalhando ao nível do *layout*, torna-se importante a existência de boas práticas de desenho. Isto porque, no posterior teste do circuito desenhado, o *layout* acarreta sempre capacidades parasitas que acabam por ter algum predomínio nos resultados das simulações pós-*layout*.

5.1 – Linhas de investigação futuras

Relativamente ao seguimento do estudo presente nesta dissertação, apresentam-se de seguida algumas sugestões que podem vir a ser exploradas no futuro. Como se conclui deste estudo, as diferenças entre o comportamento de uma tecnologia para a outra pode variar devido a inúmeros factores, alguns deles que, inevitavelmente, escapam aos modelos analíticos utilizados para o desenho de circuitos, por mais exactos que esses modelos possam ser. É então necessário dar continuidade a este estudo, fazendo mais testes e extraindo mais resultados do trabalho de simulação.

Futuramente, seria interessante um estudo que facultasse um factor de ajuste para aplicar os modelos convencionais de desenho às tecnologias actuais. Um estudo que permitisse, por exemplo, que se desenhassem portas lógicas das tecnologias de 130 nm, aplicando as técnicas tradicionais, mas com um pequeno ajuste que permitisse desenhar circuitos tão optimizados quanto o possível. Isto através de uma caracterização bastante

semelhante à feita neste trabalho, mas que possibilite inferir no final uma nova técnica de desenho, desde que devidamente fundamentada nos resultados obtidos.

Noutro contexto, propõe-se uma exploração mais efectiva dos efeitos de canal curto devido à saturação de velocidade de deriva ou mesmo à degradação da mobilidade. Num trabalho exaustivo como este, em que se trabalha com muitas tecnologias, é bastante complicado de transportar estes efeitos secundários para o modelo analítico e mais complicado se torna avaliar correctamente o impacto destes factores. Centrando-se o estudo numa das tecnologias de canal curto, vale a pena compreender e classificar a influência que certos efeitos de segunda ordem possam ter no modelo de operação do MOSFET, por forma a prever o comportamento dos circuitos que se pretendem desenhar.

Deste modo, propõe-se um estudo mais focado no modelo do transístor em si, que explore de forma detalhada o impacto dos efeitos de canal curto no desenho de circuitos. Um dos trabalhos futuros pode incluir a utilização de um modelo de correntes mais avançado e pode ser validado por exemplo com base numa análise mais minuciosa dos parâmetros dos modelos BSIM (*Berkeley Short-channel IGFET Model*). Essa análise pode ser fundamental modelizando os efeitos de segunda ordem observados com recurso a outras ferramentas, como as ferramentas de simulação matemática.

Referências

1. J. S. Kilby, “Miniaturized Electronic Circuits”, U. S. Patent 3138743, February 6, 1959.
2. R. N. Noyce, “Semiconductor Device-and-Lead Structure”, U. S. Patent 2981877, July 30, 1959.
3. Frank M. Wanlass and Chih-Tang Sah, “Nanowatt Logic Using Field-Effect Metal-Oxide Semiconductor Triodes”, International Solid-State Circuits Conference, pp. 32–33, February 1963.
4. G. E. Moore, “Cramming more components onto integrated circuits,” Electronics, pp. 114–117, Apr. 19, 1965.
5. Harold Shichman and David A. Hodges, “Modeling and Simulation of Insulated-Gate-Field-Effect Transistor Switching Circuits”, IEEE Journal of Solid-State Circuits, Vol. SC-3, No. 3, pp. 285–288, September 1968.
6. G. Baum and H. Beneking, “Drift Velocity Saturation in MOS Transistors”, IEEE Transactions on Electron Devices, Vol. ED-17, pp. 481–482, June 1970.
7. Robert H. Dennard, Fritz H. Gaensslen, L. Kuhn and H. N. Yu, “Design of micron MOS switching devices”, IEDM Technical Digest, Vol. SC-3, No. 3, pp. 168–170, December 1972.
8. Robert H. Dennard., Fritz H. Gaensslen, H. N. Yu, V. Leo Rideout, Ernest Bassous and Andre R. LeBlanc, “Design of ion-implanted MOSFET's with Very Small Physical Dimensions”, IEEE Journal of Solid-State Circuits, Vol. SC-9, No. 5, pp. 256–268, October 1974.
9. Nils Hedenstierna and Kjell O. Jeppson, “CMOS Circuit Speed and Buffer Optimization”, IEEE Transactions on Computer-Aided Design, Vol. CAD-6, No. 2, pp. 270–281, March 1987.
10. Yannis P. Tsividis, *Operation and Modeling of the MOS Transistor*, International Editions, McGraw-Hill, 1988.
11. Randall L. Geiger, Phillip E. Allen and Noel R. Strader, *VLSI Design Techniques for Analog and Digital Circuits*, McGraw Hill International Editions, 1990.
12. Srinivasa R. Vemuru and Arthur R. Thorbjornsen, “A Model for Delay Evaluation of a CMOS Inverter”, IEEE International Symposium on Circuits and Systems, May 1990.

13. Ayman I. Kayssi, Karem A. Sakallah and Timothy M. Burks, "Analytical Transient Response of CMOS Inverters", IEEE Transactions on Circuits and Systems - I, Vol. 39, pp. 42–45, January 1992.
14. H. C. Chow and W. S. Feng, "Model for Propagation Delay Evaluation of CMOS Inverter Including Input Slope Effects for Timing Verification", Electronics Letters, June 1992.
15. Neil H. E. Weste and Kamran Eshraghian, *Principles of CMOS VLSI Design: A Systems Perspective*, Second Edition, Addison-Wesley Publishing Company, 1993.
16. Kjell O. Jeppson, "Modeling the Influence of the Transistor Gain Ratio and the Input-to-Output Coupling Capacitance on the CMOS Inverter Delay", IEEE Journal of Solid-State Circuits, Vol. 29, No. 6, pp. 646–654, June 1994.
17. Nils Hedenstierna and Kjell O. Jeppson, "Comments on the Optimum CMOS Tapered Buffer Problem", IEEE Journal of Solid-State Circuits, Vol. 29, No. 6, pp. 155–158, February 1994.
18. Bijan Davari, Robert H. Dennard and Ghavam G. Shahidi, "CMOS Scaling for High Performance and Low Power – The Next Ten Years", Proceedings of the IEEE, Vol. 83, No. 4, pp. 607–618, April 1995.
19. Sung-Mo Kang and Yusuf Leblebici, *CMOS Digital Integrated Circuits: Analysis and Design*, McGraw Hill, 1996.
20. Thomas A. DeMassa and Zack Ciccone, *Digital Integrated Circuits*, John Wiley & Sons, 1996.
21. H. Wong and M. C. Poon, "Approximation of the Length of Velocity Saturation Region in MOSFET's", IEEE Transactions on Electron Devices, Vol. 44, No. 11, pp. 2033–2036, November 1997.
22. David S. Kung and Ruchir Puri, "Optimal P/N Width Ratio Selection for Standard Cell Libraries", IEEE Transactions on Computer-Aided Design, pp. 178–184, 1999.
23. J. P. Uyemura, *CMOS Logic Circuits Design*, Kluwer Academic Publishers, 1999.
24. B. T. Murphy, D. E. Haggan and W. W. Troutman, "From Circuit Miniaturization to the Scalable IC", Proceedings of the IEEE, Vol. 88, No. 5, pp. 691–703, May 2000.
25. J. S. Kilby, "The Integrated Circuit's Early History", Proceedings of the IEEE, Vol. 88, No. 1, pp. 109–111, January 2000.

26. Kaushik Roy and Sharat C. Prasad, *Low-Power CMOS VLSI Circuit Design*, Wiley-Interscience, 2000.
27. Wai-Kai Chen, *The VLSI Handbook*, CRC Press, 2000.
28. E. J. Nowak, "Maintaining the benefits of CMOS scaling when scaling bogs down", IBM Journal of Research and Development, Vol. 46, No. 2/3, pp. 169–180, March-May 2002.
29. Jan M. Rabaey, Anantha Chandrakasan and Borivoje Nikolic, *Digital Integrated Circuits: A Design Perspective*, Second Edition, Internacional Edition, Prentice Hall Electronics and VLSI Series, 2003.
30. Phaedon Avouris, J. Appenzeller, R. Martel and S. Wind, "Carbon Nanotube Electronics", Proceedings of the IEEE*,* Vol. 91, pp. 1772–1784, 2003.
31. Jean-Pierre Colinge, *Silicon-On-Insulator Technology: Materials to VLSI*, Kluwer, Third Edition, 2004
32. Phaedon Avouris, "Supertubes", IEEE Spectrum, August 2004.
33. R. Jacob Baker, *CMOS Circuit Design, Layout, and Simulation*, Second Edition, Wiley-IEEE Press, 2004
34. Mohamed Elgebaly, "Energy Efficient Design for Deep Sub-micron CMOS VLSIs", PhD Thesis, University of Waterloo, 2005.
35. Thomas Skotnicki, James A. Hutchby, Tsu-Jae King, H.-S. Philip Wong and Frederic Boeuf, "The End of CMOS Scaling: Toward the Introduction of New Materials and Structural Changes to Improve MOSFET Performance", IEEE Circuits & Devices Magazine, pp. 16–26, January/February 2005.
36. Boris Andreev, Edward L. Titlebaum e Eby G. Friedman, "Sizing CMOS Inverters with Miller Effect and Threshold Voltage Variations", Journal of Circuits and Computers, Vol. 15, No. 3 (2006), pp. 437–454, March 2006.
37. T. C. Chen, "Where Si-CMOS is Going: Trend Hype vs Real", The International Solid-State Circuits Conference, 2006.
38. El-Sayed A. M. Hasaneen, Mohamed A. A. Wahab and Osama N. A. Esmali, "MOSFET Scaling to Nanometer Regimes", IEEE Journal of Solid-State Circuits, pp. 261–264, December 2007.
39. Mark T. Bohr, Robert S. Chau, Tahir Ghani and Kaizad Mistry, "The High-K Solution", IEEE Spectrum, October 2007.